软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(8):3637−3654 [doi: 10.13328/j.cnki.jos.007354] [CSTR: 32375.14.jos.007354]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

因果时空语义驱动的深度强化学习抽象建模方法*

田丽丽[1,3,4], 杜德慧[2,3,4], 聂基辉[2,4], 陈逸康[2,4], 李荣达[2,4]

[1](华东师范大学 计算机科学与技术学院, 上海 200062)

[2](华东师范大学 软件工程学院, 上海 200062)

[3](华东师范大学 智能教育研究院, 上海 200062)

[4](上海市高可信计算重点实验室 (华东师范大学), 上海 200062)

通信作者: 杜德慧, E-mail: dhdu@sei.ecnu.edu.cn

摘　要: 随着智能信息物理融合系统 (intelligent cyber-physical system, ICPS) 的快速发展, 智能技术在感知、决策、规控等方面的应用日益广泛. 其中, 深度强化学习因其在处理复杂的动态环境方面的高效性, 已被广泛用于 ICPS 的控制组件中. 然而, 由于运行环境的开放性和 ICPS 系统的复杂性, 深度强化学习在学习过程中需要对复杂多变的状态空间进行探索, 这极易导致决策生成时效率低下和泛化性不足等问题. 目前对于该问题的常见解决方法是将大规模的细粒度马尔可夫决策过程 (Markov decision process, MDP) 抽象为小规模的粗粒度马尔可夫决策过程, 从而简化模型的计算复杂度并提高求解效率. 但这些方法尚未考虑如何保证原状态的时空语义信息、聚类抽象的系统空间和真实系统空间之间的语义一致性问题. 针对以上问题, 提出基于因果时空语义的深度强化学习抽象建模方法. 首先, 提出反映时间和空间价值变化分布的因果时空语义, 并在此基础上对状态进行双阶段语义抽象以构建深度强化学习过程的抽象马尔可夫模型; 其次, 结合抽象优化技术对抽象模型进行调优, 以减少抽象状态与相应具体状态之间的语义误差; 最后, 结合车道保持、自适应巡航、交叉路口会车等案例进行了大量的实验, 并使用验证器 PRISM 对模型进行评估分析, 结果表明所提出的抽象建模技术在模型的抽象表达能力、准确性及语义等价性方面具有较好的效果.

关键词: 深度强化学习; 抽象建模; 因果时空语义; 智能信息物理融合系统 (ICPS); 马尔可夫决策过程 (MDP)

中图法分类号: TP18

中文引用格式: 田丽丽, 杜德慧, 聂基辉, 陈逸康, 李荣达. 因果时空语义驱动的深度强化学习抽象建模方法. 软件学报, 2025, 36(8): 3637–3654. http://www.jos.org.cn/1000-9825/7354.htm

英文引用格式: Tian LL, Du DH, Nie JH, Chen YK, Li YD. Causal-spatiotemporal-semantics-driven Abstraction Modeling Method for Deep Reinforcement Learning. Ruan Jian Xue Bao/Journal of Software, 2025, 36(8): 3637–3654 (in Chinese). http://www.jos.org.cn/1000-9825/7354.htm

Causal-spatiotemporal-semantics-driven Abstraction Modeling Method for Deep Reinforcement Learning

TIAN Li-Li[1,3,4], DU De-Hui[2,3,4], NIE Ji-Hui[2,4], CHEN Yi-Kang[2,4], LI Ying-Da[2,4]

[1](School of Computer Science and Technology, East China Normal University, Shanghai 200062, China)

[2](Software Engineering Institute, East China Normal University, Shanghai 200062, China)

[3](Institute of AI Education, East China Normal University, Shanghai 200062, China)

[4](Shanghai Key Laboratory of Trustworthy Computing (East China Normal University), Shanghai 200062, China)

Abstract: With the rapid advancement of intelligent cyber-physical system (ICPS), intelligent technologies are increasingly utilized in components such as perception, decision-making, and control. Among these, deep reinforcement learning (DRL) has gained wide

* 本文由“形式化方法与应用”专题特约编辑陈明帅研究员、田聪教授、熊英飞副教授推荐.
收稿时间: 2024-08-26; 修改时间: 2024-10-14; 采用时间: 2024-11-26; jos 在线出版时间: 2024-12-10
CNKI 网络首发时间: 2025-04-17

application in ICPS control components due to its effectiveness in managing complex and dynamic environments. However, the openness of the operating environment and the inherent complexity of ICPS necessitate the exploration of highly dynamic state spaces during the learning process. This often results in inefficiencies and poor generalization in decision-making. A common approach to address these issues is to abstract large-scale, fine-grained Markov decision processes (MDPs) into smaller-scale, coarse-grained MDPs, thus reducing computational complexity and enhancing solution efficiency. Nonetheless, existing methods fail to adequately ensure consistency between the spatiotemporal semantics of the original states, the abstracted system space, and the real system space. To address these challenges, this study proposes a causal spatiotemporal semantic-driven abstraction modeling method for deep reinforcement learning. First, causal spatiotemporal semantics are introduced to capture the distribution of value changes across time and space. Based on these semantics, a two-stage semantic abstraction process is applied to the states, constructing an abstract MDP model for the deep reinforcement learning process. Subsequently, abstraction optimization techniques are employed to fine-tune the abstract model, minimizing semantic discrepancies between the abstract states and their corresponding detailed states. Finally, extensive experiments are conducted on scenarios including lane-keeping, adaptive cruise control, and intersection crossing. The proposed model is evaluated and analyzed using the PRISM verifier. The results indicate that the proposed abstraction modeling technique demonstrates superior performance in abstraction expressiveness, accuracy, and semantic equivalence.

Key words: deep reinforcement learning (DRL); abstraction modeling; causal spatiotemporal semantics; intelligent cyber-physical system (ICPS); Markov decision process (MDP)

1 引 言

信息物理融合系统 (cyber-physical system, CPS)[1] 是集计算、通信和物理环境的复杂系统, 具有混成性、复杂性和实时性等特性, 通常运行在开放、不确定性环境中, 例如自动驾驶、智慧医疗、智慧城市、智能交通等环境. CPS 系统的控制部件是系统的核心, 需要根据感知环境得到的信息进行智能控制, 常见的控制器包括模型预测控制 (model predictive control, MPC)、比例积分微分 (proportional-integral-derivative, PID) 控制、线性二次调节器 (linear quadratic regulator, LQR) 等. 以自动驾驶的自适应巡航控制系统 (adaptive cruise control, ACC) 为例, 模型预测控制器接收智能体与前车的相对距离、用户设置的巡航速度、自车速度和与前车的相对速度等输入信息, 然后将加速度等控制信号传递给自车执行器, 执行相应的动作. 传统的 MPC 控制器通过预测有限时间内两车的运动, 每个时间步都生成控制命令, 以在保持与前车安全距离的同时, 实现用户设置的巡航速度跟踪目标.

近年来, 机器学习技术在 CPS 系统中得到了广泛应用. 例如, 在自动驾驶领域, 感知部件可以借助卷积神经网络 (convolutional neural network, CNN) 网络来识别路牌、行人、障碍物等; 决策部件采用强化学习等技术实现智能决策. 此类融合智能组件来实现系统功能运作的系统称为智能 CPS (intelligent CPS, ICPS)[1]. 深度强化学习 (deep reinforcement learning, DRL)[2] 结合了深度学习的近似能力与强化学习的决策能力, 能够处理高维、连续、复杂的状态和动作空间问题. 它不仅能够基于原始输入数据自动提取有意义的特征, 而且可以在无监督的情况下通过与环境的交互, 学习到策略模型. DRL 已经被广泛用于非线性、随机和高度不确定性的系统, 为其提供控制优化策略. 虽然深度强化学习在 ICPS 中取得了显著的成就, 但它仍面临一系列问题, 如规模性问题[3], 即智能体在每个时间步接收环境的状态信息并进行一次决策, 需要智能体需要在细粒度的状态空间和决策时间内进行操作, 从而引发了强化学习任务的规模性问题. 具体而言, 大规模的状态空间导致了状态空间探索效率低以及奖励稀疏[4] 等问题. 其次, 长期决策的过程导致了轨迹空间规模庞大, 使强化学习的目标函数难以优化[5], 进而导致学习效率低、泛化能力弱以及算法稳定性差等问题.

解决强化学习的规模性问题的一种有效方法是使用抽象建模技术将大规模、细粒度的马尔可夫决策过程抽象为小规模粗粒度的马尔可夫决策过程[6], 从而将大规模复杂的决策任务抽象为小规模简单的决策任务, 减小状态空间以及轨迹空间的规模. 现有的强化学习的抽象技术主要分为 3 类: 状态抽象、动作抽象和状态-动作联合抽象. 状态抽象是空间尺度上的抽象, 即利用状态抽象函数将大规模状态空间抽象为小规模的状态空间[7]. 动作抽象是时间尺度上的抽象, 即利用抽象动作策略将每单步决策的智能体决策过程抽象为每多步决策的智能体决策过程[8]. 状态-动作联合抽象是同时在状态空间尺度和动作时间尺度上做联合抽象, 旨在解决强化学习的规模性问题[9].

一个有效的抽象需要在降低复杂性的同时, 尽可能保证抽象后的最优性与原问题保持一致[10]. 然而, 将抽象建模技术应用于强化学习的工作虽然已经取得了初步的研究成果, 但是在保留原问题最优性上仍面临挑战, 特别是空间语义、时间语义以及概率语义信息未能被完整保留. 这种信息丢失可能导致最优策略的丢失、泛化能力减弱和策略鲁棒性降低等风险, 例如自动驾驶车辆在城市道路行驶面临多种复杂的交通状况 (行人穿越马路、车辆并线等), 抽象建模中通常将城市道路简化为网格状的空间表示, 若空间语义丢失, 即车辆位置被粗略的表示为某个网格中的点, 而不是精确的位置, 导致车辆在复杂交通中需要切换车道时, 丢失空间语义信息的抽象模型无法准确判断何时应该并线以避免与其他车辆相撞, 导致车辆采取次优的驾驶策略, 甚至引发交通事故. 因此, 亟需探索新的抽象技术来应对这一挑战, 使抽象后的强化学习过程最大化地适用于 ICPS 控制器的决策生成, 并确保系统的安全性.

针对上述问题, 本文基于因果关系推理理论提出了一种基于因果时空语义对状态空间进行分层抽象的方法. 首先, 因果时空语义兼顾状态的时间信息、空间信息和概率信息, 从状态的本质出发, 对具体状态语义的每个维度进行了第 1 阶段的抽象, 将复杂的状态空间进行分解和简化, 以便于对状态空间进行有效的分析和理解. 其次, 提出度量时间和空间变化的价值分布的时空价值矩阵, 并基于时空价值矩阵进行聚类抽象, 实现模型的第 2 阶段抽象. 该方法提高了模型抽象的程度, 使得抽象后的 MDP[5] 模型更加精简. 此外, 由于状态和动作是密切相关的, 本文提出在第 2 阶段抽象过程中需要兼顾动作抽象, 以实现状态-动作联合抽象, 达到最优的抽象效果. 最后, 结合 ICPS 的典型案例进行了车道保持、自适应巡航、交叉路口会车等多组对比实验分析, 实验结果表明基于因果时空语义的双层抽象方法具有较好的准确性和简洁性.

本文第 1 节介绍问题提出的背景和重要意义. 第 2 节介绍本文所需的背景知识, 包括智能信息物理融合系统、基于强化学习的控制生成、MDP 以及抽象 MDP. 第 3 节介绍抽象建模方法的研究现状和存在的问题. 第 4 节介绍如何基于因果时空语义构建抽象模型. 第 5 节通过 3 个案例对本文的方法进行实验, 并对实验结果进行分析. 最后一节总结全文.

2 背景知识

本节将详细介绍 ICPS、基于深度强化学习的控制器和抽象技术的相关概念及基本知识.

2.1 智能信息物理融合系统控制器 (ICPS)

智能信息物理融合系统控制器 (ICPS) 是在信息物理融合系统中融入人工智能技术, 帮助实现智能感知、智能决策. 如后文图 1 所示, ICPS 主要由以下 4 个部分构成: 基于机器学习的控制器、执行器、物理环境以及传感器. 在 ICPS 中, 传感器和控制器分别负责感知和决策. 传感器在 ICPS 中扮演着感知器官的角色, 负责采样物理世界的连续状态, 并将其转换为离散的信号, 在每个时间步 t 后输出新的系统状态 s_{t+1}. 控制器则利用传感器接收到的系统状态 s_t 及外部输入信号 i_t, 根据所学策略输出控制信号 c_t, 以引导执行器的行动, 从而实现 ICPS 的控制功能. 执行器根据控制命令 c_t 调整智能体在实际物理环境中的行为, 使系统达到预期状态. 实际物理环境是 ICPS 中的关键组成部分, 通过 ICPS 的非线性连续动力学模型 M_{env}, 可以计算当前系统状态 s_t 和执行器的输出 c_t, 从而得到下一个系统状态 s_{t+1}.

2.2 基于深度强化学习的控制生成

深度强化学习将强化学习 (reinforcement learning, RL) 与深度学习 (deep learning) 相结合, 用于训练智能体在复杂环境中进行决策. 具体而言, 深度强化学习通过深度神经网络 (deep neural network, DNN) 来解决复杂的非线性问题, 使智能体具有处理高维数据的能力, 从而使 ICPS 能够在复杂环境中实现自主决策功能. 如图 2 所示, 在深度强化学习的框架中, 智能体通过与环境交互来学习最优策略. 每个时间步骤中, 智能体接收传感器获取的环境状态数据 s_t, 并使用 DNN 对这些数据进行处理和评估, 随后输出当前环境下的最优动作 a_t. 智能体的决策受到探索策略 (如 ϵ-贪婪策略) 的影响, 该策略用于对未知环境的探索与对已知信息的利用. 在执行动作后, 智能体会根据环境反馈获得奖励信号 r_t, 以衡量决策的效果. 智能体通过这些奖励来调整决策过程, 优化未来的策略, 以获得更高的累积奖励. 通过这种方式, 深度强化学习能够有效地提升智能体的决策能力, 从而实现长期目标.

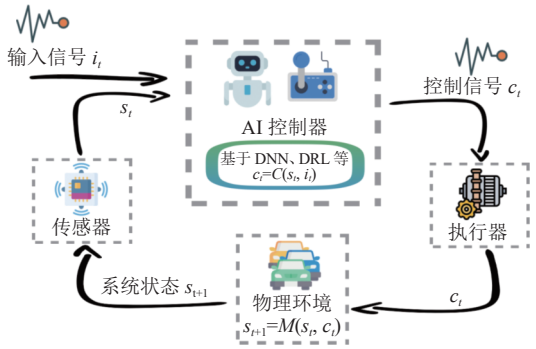


图 1 智能信息物理融合系统

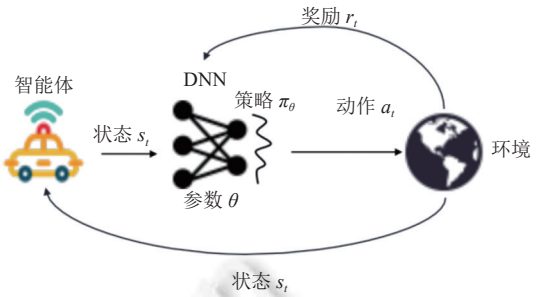


图 2 深度强化学习

2.3 马尔可夫决策过程

深度强化学习可以通过马尔可夫决策过程 (MDP)[11]进行形式化描述.

定义 1. 马尔可夫决策过程. 由一个五元组构成 $M=(S,A,R,P,\gamma)$, 其中, S 表示有限非空状态集合, A 表示有限非空动作集合, $P:S\times A\times S\to[0,1]$ 表示迁移概率函数, 对于 $s\in S$, $a\in A$, $\Sigma_{s'\in S}P(s,a,s')=1$. $R:S\times A\to\mathcal{R}$ 为当前状态-动作对分配奖励值, 而 $\gamma\in(0,1)$ 表示折扣因子. 折扣因子 γ 决定了即时奖励对未来奖励的重要性. 较大的 γ 将使智能体从长期奖励中学习.

在 MDP 中, 策略 $\pi:S\to A$ 将状态集合映射到动作集合. $\pi(s)$ 表示在状态 s 下应采取的动作, 马尔可夫决策过程 M 描述了系统初始状态在离散时间步中的演变. 在 DRL 中, 策略 π 与环境交互所产生的状态转移和即时奖励构成了学习过程的核心基础, 而策略的评估和优化则依赖于状态价值函数和动作价值函数. 状态价值函数如公式 (1) 所示, $V(s)$ 表示在状态 s 下, 遵循某一策略 π 后智能体能获得的累计奖励期望值. 该值越高表示从该状态出发, 在按照特定策略行动时可以获得更高的长期奖励. 动作价值函数如公式 (2) 所示, $Q(s,a)$ 表示在状态 s 下, 采取动作 a 并遵循某一策略 π 后能获得的长期期望回报. 动作价值函数用于评估特定状态下采取不同动作的价值, 从而帮助智能体最大化期望回报.

$$V^{\pi}(s)=E_{\pi}[G_t|S_t=s] \tag{1}$$

$$Q^{\pi}(s,a)=E_{\pi}[G_t|S_t=s,A_t=a] \tag{2}$$

其中, G_t 是从时间 t 开始的回报, $E_{\pi}[\cdot]$ 表示策略 π 下的期望值.

状态价值函数和动作价值函数之间具有密切的联系. 具体而言, 状态价值函数可以通过动作价值函数来计算, 反之亦然. 如公式 (3) 所示, 状态价值函数可被视为特定状态下所有可能动作的期望动作价值; 公式 (4) 所示, 若已知每个动作的动作价值, 则可以选择最佳动作来确定该状态的最优价值.

$$V^{\pi}(s)=\sum_a\pi(a|s)[Q^{\pi}(s,a)] \tag{3}$$

$$Q^{\pi}(s,a)=R(s,a)+\gamma\sum_{s'}P(s'|s,a)V^{\pi}(s') \tag{4}$$

其中, $R(s,a)$ 是执行动作 a 时从状态 s 获得的即时奖励, γ 是折扣因子, 表示未来奖励的当前价值, $P(s'|s,a)$ 是从状态 s 执行动作 a 转移到状态 s' 的概率. 深度强化学习通过这些函数引导智能体在给定环境中学习最优策略, 即使面对任何状态也能够选择最佳动作以实现长期奖励的最大化.

2.4 抽象马尔可夫决策过程

对于现实世界中的 ICPS 而言, 它们所面临的 MDP 模型状态空间通常庞大而复杂, 这使得在真实 MDP 中基于深度强化学习进行决策生成变得具有挑战性. 为了解决这一问题, 抽象 MDP 通过将真实状态映射为简化状态, 将原本复杂庞大的状态空间转化为更小、更简单且易于处理的抽象状态空间, 同时尽可能保持策略的最优性. 这里的最优性指的是, 在模型抽象后, 所学到的策略仍能确保在原始问题上实现相同的最优累积奖励, 即抽象后的

MDP 能够保持与原始 MDP 中相同的最优策略[7].

定义 2. 抽象马尔可夫模型. 设 $M=(S,A,R,P,\gamma)$ 表示真实的 MDP, 其对应的抽象 MDP 表示为 $\bar{M}=(\bar{S},\bar{A},\bar{P},\bar{R},\gamma)$, $\Phi: S \to \bar{S}$ 表示状态抽象函数, $\Phi(s)\in\bar{S}$ 为抽象状态, 其逆映射表示为 $\Phi^{-1}(\bar{s})$, 其中 $\bar{s}\in\bar{S}$. $\bar{S}$ 是与抽象函数 Φ 对应的抽象状态集, $\Psi: A \to \bar{A}$; $\Psi(a)\in\bar{A}$ 表示动作抽象函数, 其逆映射表示为 $\Psi^{-1}(\bar{a})$, $\bar{a}\in\bar{A}$. $\bar{A}$ 是与动作抽象函数 Ψ 对应的抽象动作集.

根据抽象 MDP, 奖励函数和状态转移可以定义为:

$$\bar{R}(\bar{s},\bar{a})=\sum_{s\in\Phi^{-1}(\bar{s}),\, a\in\Psi^{-1}(\bar{a})} w(s)v(a)R(s,a) \tag{5}$$

$$\bar{P}^{\bar{a}}_{\bar{s}\bar{s}'}=\sum_{s\in\Phi^{-1}(\bar{s}),\, a\in\Psi^{-1}(\bar{a})}\sum_{s'\in\Phi^{-1}(\bar{s}')} w(s)v(a)P^{a}_{ss'} \tag{6}$$

其中, $w: S \to [0,1]$, $\bar{s}\in\bar{S}$, $\sum_{s\in\Phi^{-1}(\bar{s})} w(s)=1$ 表示状态的权重函数, $v: A\to[0,1]$, $\bar{a}\in\bar{A}$, $\sum_{a\in\Psi^{-1}(\bar{a})} v(a)=1$ 表示动作的权重函数. $\bar{R}(\bar{s},\bar{a})$ 表示在执行抽象动作 $\bar{a}$ 后从抽象状态 $\bar{s}$ 转移到 $\bar{s}'$ 的即时奖励, $\bar{P}^{\bar{a}}_{\bar{s}\bar{s}'}$ 表示在执行抽象动作 $\bar{a}$ 后从抽象状态 $\bar{s}$ 转移到 $\bar{s}'$ 的概率.

3 相关工作

为解决强化学习由于需探索复杂多变的状态空间而造成决策生成时效率低下泛化性差等问题[6], 我们发现已有工作通过将复杂的大规模细粒度 MDP 抽象为小规模的粗粒度 MDP, 以缩小状态空间的规模, 降低计算负担. 在深度强化学习的研究和应用中, 根据抽象对象的不同, 抽象方法主要划分为 3 类: 动作抽象、状态抽象及状态-动作对抽象, 下面将分别对这 3 大类方法进行总结.

动作抽象是通过在时间维度上减少智能体决策所需步骤, 优化决策过程, 并促进智能体进行长期的策略学习. 目前的研究工作中将动作抽象引申出了宏动作 (options)、子任务 (subtasks)、技能 (skill)[5]、子目标 (subgoal)[12] 以及子策略 (sub-policy)[13] 等概念. 有研究通过减少必须进行的决策数量来解决需要长期决策过程的问题[5]. 除此之外, 为了提高学习效率, 动作抽象通过为智能体设定子目标并给予内在奖励来解决奖励稀缺环境下的问题[8]. 同时, 还通过设置与特定任务无关的子目标来增强模型在不同任务之间的泛化能力[14]. 然而, 由于庞大状态空间导致巨大抽象动作空间存在困难, 基于抽象动作的强化学习优化仍面临挑战.

状态抽象则是在空间尺度上的抽象, 其目标是将环境中的原始大规模状态空间抽象为小规模状态空间, 从而降低强化学习算法的样本复杂度以及探索难度. 然而, 由于信息丢失、抽象非一致性等问题所生成的策略无法保证与真实 MDP 所生成策略完全一致的最优性[15]. 因此, 研究人员提出了多种度量状态相似性的方法, 以期望经过抽象后仍能保持原决策过程最优性. 例如, Castro[7] 使用深度神经网络度量 MDP 中的状态相似性; Taïga 等人[16] 提出了近似 MDP 同胚理论, 专注于合并具有相似迁移概率和奖励的状态; Taylor 等人[17] 则提出了一个宽松的互模拟等价度量, 将其与近似 MDP 同胚理论结合, 以达到状态抽象的目的. Junges 等人[3] 将 MDP 视为层次结构, 将状态空间划分为宏观级别和子级别, 将不可再划分的子级别视为约束, 不断对约束进行划分, 分析模型的不确定性, 以解决状态空间爆炸问题. 清华大学 Feng 等人[18] 提出了 DenseRL 方法, 通过对现有自动驾驶数据集中的安全关键场景的 MDP 进行编辑, 鼓励 ADS 在这些安全关键场景下进行测试, 展示了状态抽象技术在 ICPS 中的应用潜力. 但是, 状态聚类抽象研究存在一个难点即如何保证原状态的时空语义信息.

状态-动作联合抽象同时在状态空间尺度和动作时间尺度上进行抽象, 以结合状态抽象和动作抽象的优点. Abel 等人[19] 定义了 4 种不同的状态-动作对的抽象类别, 并对原始问题提供次优解决方案. Song 等人[20] 提出 SIEGE, 将系统状态规约为性质语义, 以性质语义联合动作进行抽象, 但是规约性质缺乏对概率语义的描述, 会损失状态的概率语义. Guo 等人[21] 提出测地度量对系统状态进行聚类抽象, 考虑抽象模型能否保留原始真实模型的最优性, 但是测地度量更强调状态的空间维度相似度. 此外, 还可以利用神经网络进行表征学习的方法实现状态动作对抽象, 如 FuN[22] 将复杂的高维系统空间映射至低维的抽象表示, 以此来减少状态-动作对的维度. 从本质上来看, 状态-动作对抽象是一种通过离散化状态和动作, 并求其笛卡尔积后实施的抽象方法. 然而, 这一过程常导致状态空

间进一步增大. 尽管采用神经网络进行表征学习的方式可以在某种程度上压缩状态空间, 但目前还不清楚上述聚类抽象的系统空间和真实系统空间是否具有语义一致性.

此外, 因果推理是一门研究因果关系及其在各种复杂系统和现象中的作用与影响的科学[23]. 它通过分析变量之间的因果联系, 揭示事物发展的内在机制和规律, 从而为决策和优化提供理论依据和方法支持. 因此, 本文提出基于因果时空语义的双阶段抽象方法, 从 ICPS 状态价值语义、动作价值语义以及迁移概率语义等方面进行剖析, 刻画系统不同时间空间下的价值. 通过将这些涵盖的语义作为抽象输入, 进而实现保持抽象模型与真实模型的语义等价性, 并缩小模型规模.

4 基于因果时空语义的抽象模型构建

本节将探讨模型抽象的核心问题——如何度量不同状态之间的相似度, 并据此判断它们是否可归入为同一抽象状态. 解决该问题对于构建 ICPS 场景中的抽象 MDP 模型至关重要. 为此, 本文提出了一种基于因果时空语义的深度强化学习抽象建模方法, 该方法在保持真实马尔可夫决策过程语义信息的同时, 实现状态的有效抽象. 具体来说, 第 1 阶段从具体状态的各个特征出发, 依据特征间的因果关系, 确定特征的语义计算方法. 通过衡量不同语义粒度划分对语义特征抽象的影响, 实现对状态特征的精确抽象. 第 2 阶段, 借助时空语义信息, 构建了时空语义度量, 并采用 (ε,d)-抽象方法, 构建场景的抽象 MDP 模型. 图 3 展示了基于因果时空语义的深度强化学习抽象建模的方法框架, 主要包含以下 3 个部分.

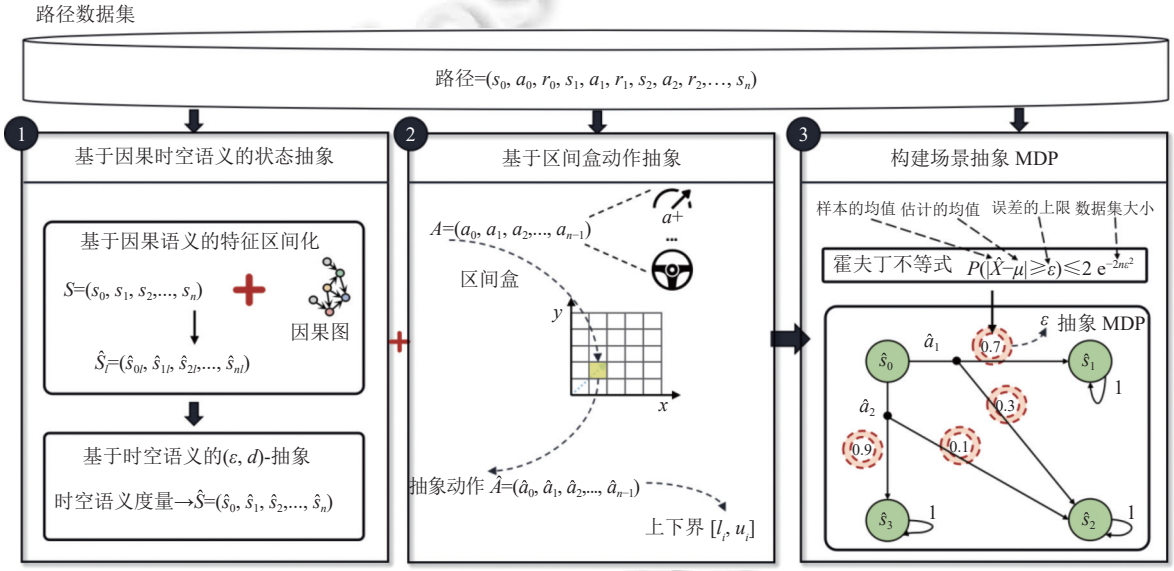


图 3 基于因果时空语义的抽象方法框架

(1) 基于因果时空语义的状态抽象: 因果时空语义涵盖状态特征间的因果关系以及状态之间的时空关系. 根据状态特征间的因果语义, 实现对状态特征的压缩, 并进行区间化处理. 时空语义引入了价值信息、时空信息和概率信息, 通过时空语义度量评估状态之间的相似程度, 从而进行状态抽象. 时空价值语义能够更全面地捕捉状态之间的关联性, 提高抽象模型的准确性和实用性.

(2) 基于区间盒的动作抽象: 通过区间盒抽象, 将连续动作空间等距离地划分为单位区间, 并根据具体环境和问题的特性调整抽象粒度, 以提升强化学习算法对动作选择问题的处理能力, 同时降低计算复杂度.

(3) 构建场景的抽象 MDP: 收集安全路径数据集, 提出迁移空间的概念, 并运用霍夫丁不等式设计迁移概率计算公式, 以提高迁移概率的准确性. 同时, 结合抽象状态和抽象动作, 构建场景的抽象 MDP 模型.

4.1 基于因果时空语义的状态抽象

在构建抽象模型时, 模型需要在简洁性与准确性之间找到一个恰当的平衡点. 简洁性要求有效控制状态数量, 而准确性则要求减少抽象状态与具体状态之间的误差. 因此, 抽象建模方法既要保证模型简洁易用, 又要确保其高度实用和准确, 使得通过语义抽象得到的模型能够真实地反映系统的实际情况. 本文提出了基于因果时空语义构建抽象模型, 以确保所得到的模型既简洁又准确. 因果时空语义包括状态特征间的因果语义和状态之间的时空语义, 为系统行为提供了一个全面的多维度理解和表征方法. 通过因果关系映射, 本文将复杂的高维状态空间抽象到更加抽象的状态空间表示, 并通过对状态之间时空关系进度量来缩减庞大的状态空间. 该方法能够准确反映出系统需满足的规约性质, 并展示不同需求下的系统行为特征, 不仅能够更有效地捕捉系统决策核心含义, 还提高了基于抽象模型进行决策生成的效率.

4.1.1 基于因果语义的特征区间化抽象

在 ICPS 中, 传感器数据的高维复杂性给决策带来了挑战. 由于实际应用于决策的传感器数据相对稀缺, 因此需要对这些数据进行有效地处理. 单独一个特征并不能提供足够的信息支持决策过程, 而不同维度的组合可能会为决策带来更加丰富和全面的场景理解. 此外, 在处理传感器数据时还需要考虑到不同类型、不同精度、甚至是不同时间尺度下数据之间可能存在着潜在联系与影响. 因此, 发现特征间的关系并进行组合以实现状态维度压缩变得至关重要. 本文结合自动驾驶的数据集, 应用现有的因果发现算法, 例如, PC (Peter-Clark algorithm) 算法[24]和 FCI (fast causal inference) 算法[25]构建因果图. 借助因果图识别出不同状态特征之间的因果关系, 并根据这些关系设计特征抽象映射函数, 将原始的高维度的状态空间映射到一个更加简洁、更易于处理的抽象状态空间.

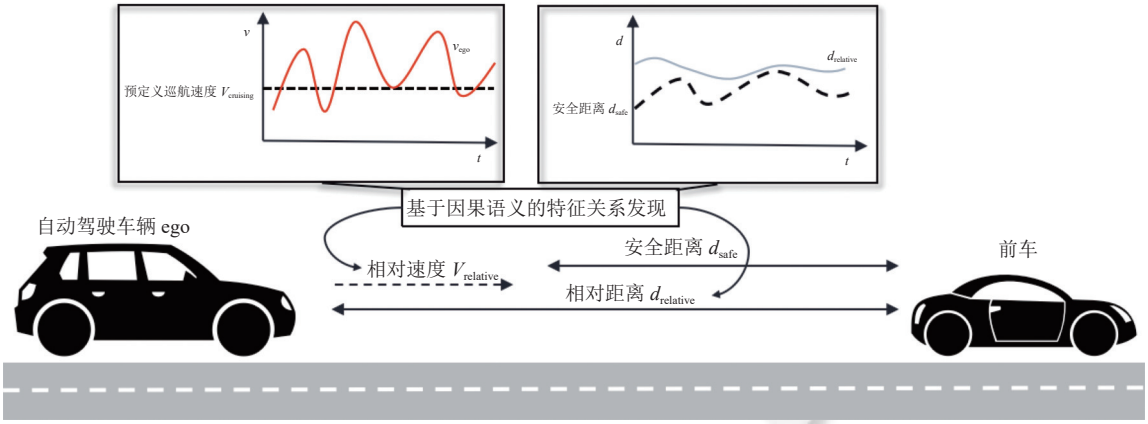
定义 3. 基于因果发现的因果关系获取. 因果推理任务可以获取所有被观测变量间的因果关系[23], 输入为观测到的变量, 输出为一种能够表示因果关系的因果图 G, 其表示形式为一个有向无环图 $G=(V,E)$, 其中 V 是节点集合, 由原始数据中观察到的变量组成, E 为边的集合, 其中, 边表示因果关系, 并且图中没有有向环, 即不存在从一个节点出发经过有向边回到自身的路径.

在本文中, 一个被观测的变量就是 MDP 中的一个状态 $s_i \in S$. 在一些假设下, 给定一个数据集 D, 称还原真实因果图 G 的任务为因果发现. 真实因果图 G 要求满足一致性和可识别性假设. 一致性指通过因果图导出的概率分布中蕴含的独立性与 D 中的一致. 可识别性指因果图中的所有边的方向都被确定. 为了满足一致性, 因果发现算法引入因果忠实性假设, 该假设要求因果图导出的概率分布的独立性能够蕴含因果图本身所表达的独立性, 这使得这些算法能够重点关注图空间上的非参数性质.

本文使用的因果发现算法为满足因果忠实性和因果无环性假设的 PC 算法[24]和 FCI 算法[25]. 该研究中, 我们首先使用 PC 和 FCI 算法对观测数据进行初步的因果关系挖掘, 以应对数据间复杂的线性和非线性关系, 并排除观测特征中非因果关系的观测特征, 在此基础上, 使用互信息方法对初步挖掘到的因果关系进行确定[26,27], 最后得到的结果即为我们想要的因果关系. 该方法的输入为观测数据也即 ICPS 数据集特征变量 $s=(d_1,d_2,d_3,\ldots,d_n)$, s 为真实状态且具有 n 个特征, 输出为不同特征状态之间的因果关系, 以 ICPS 数据集为输入得到的因果图为 $G=(s,E)$, s 为状态集合, E 为状态集合之间的因果关系.

定义 4. 基于因果语义的特征抽象. 设 $s=(d_1,d_2,d_3,\ldots,d_n)$ 表示真实状态, s 具有 n 个特征. Θ 表示语义映射函数, $\theta=\Theta(d_i,\ldots,d_k)$ 表示存在因果关系的 $d_i,\ldots,d_k$ 特征经过 Θ 映射的具体语义值 θ. $\bar{s}=(\theta_1,\ldots,\theta_j)$ 表示经过基于语义的特征抽象后的抽象状态.

如图 4 所示, 我们对基于因果语义的特征区间化抽象进行举例说明, 以自适应巡航控制为例, 其目标是追寻前车并保持安全距离. 自动驾驶车辆状态 s 用多维向量 $(v,acc,x,y,\ldots)$ 表示, 分别表示车辆速度 v、加速度 acc 及空间坐标 (x,y) 等. 通过对因果关系进行挖掘并因果图进行分析, 可以实现基于因果语义的抽象. 基于因果语义的抽象将具体状态 $s=(v,acc,x,y,\ldots)$ 简化为表示 $d=(rel_{\text{velocity}},rel_{\text{angle}},rel_{\text{distance}},\ldots)$, 其中 rel_{velocity}、rel_{angle}、rel_{distance} 分别代表相对速度、相对角度和相对距离. 通过基于因果发现的因果关系获取步骤, 我们得知角度和速度之间不存在因果关系, 故无需特征映射. 基于因果关系筛选的特征抽象方法保留了 ICPS 所需的关键信息, 有效地减少了状态空间的复杂性, 并为后续决策和控制提供更高效且可解释性更强的状态表示.



速度语义 θ^{pv}: $V_{relative}=V_{ego}-V_{cruising}$

距离语义 θ^{pd}: $d_{relative}=Position_{front}-Position_{ego}>d_{safe}$

图 4 基于因果语义的特征区间化抽象示意图

经过因果关系映射后, 发现语义特征 θ_i 的范围并不一致, 且表现形式仍然是连续范围内的具体数值. 为了实现基于特征的区间化抽象, 需要将语义值归一化到统一的范围, 并以此为基础进行离散化操作. 这个过程可以被视作将一个多维空间划分为若干个区间, 以确保每个特征取值都在可控范围内.

假设状态 $S=(\theta_1,\theta_2,\theta_3,\ldots,\theta_J)$ 拥有 J 维语义空间, 需要将 J 维空间划分为 $\prod_{j=1}^{J}K_j$ 个段, 每个维度上有 K_j 个区间, 表示为 $d_i^j=\left[l_i^j,u_i^j\right]$, 其中, d_i^j 是第 j 维上的第 i 个区间, l_i^j 和 u_i^j 是该区间的下界和上界, $1\leqslant i\leqslant K_j$. 基于区间化的划分可以在整个空间中建立一个有序的结构, 这样后续处理起来更方便和高效. 但是需要确保这种划分是合理的, 因此需要将空间划分问题转化为了优化问题, 具体表述如下:

$$\begin{cases}\max(u_i^j-l_i^j)\\ \text{s.t.}\begin{cases}d_{\text{MIN}}^j\leqslant u_i^j-l_i^j\leqslant d_{\text{MAX}}^j\\ \left|\hat{s}_i^j\right|\geqslant n_{\text{MIN}}^j\\ MEAN\left\{\theta_s^j-E\left[\hat{\theta}_s^j\right]\right\}<e_{\text{MEAN}}^j\\ MAX\left(\hat{\theta}_s^j\right)-E\left[\hat{\theta}_s^j\right]<e_{\text{MAX}}^j\end{cases}\end{cases}\tag{7}$$

其中, d_{MIN}^j 和 d_{MAX}^j 分别是第 j 个语义维度上区间的最小长度和最大长度, $\hat{s}_i^j=\left\{s|\theta_s^j\in d_i^j\right\}$ 是语义值 θ_s^j 落在区间 d_i^j 内的具体状态的集合, n_{MIN}^j 是第 j 个维度区间中具体状态的最小数量, e_{MEAN}^j 和 e_{MAX}^j 是第 j 个维度上抽象误差的预定义平均值和最大误差, $MEAN$ 指的是均值函数, MAX 指的是最大值函数. 这些公式确保每个区间包含足够的具体状态, 同时保持较低的抽象误差.

算法 1 描述了基于因果语义的特征区间化抽象算法, 该方法将复杂的具体状态集合 S 转化为抽象且具有丰富语义信息的抽象空间 $\hat{S}$. 在这个过程中, 使用因果语义映射函数 Θ 并遵循区间最大长度 d_{MAX} 和最小长度 d_{MIN}、区间内最小具体状态数 n_{MIN} 以及期望误差范围 e_{MEAN} 和状态压缩指标 r_d 等约束条件. 有学者提出在评估状态压缩效果时需要考虑参数 r_d, 其目标是在保证语义信息不受损害前提下有效减少状态数量[20]. 通过迭代优化以及语义损失衡量[20]后, 该算法在不牺牲重要语义信息情况下将原始状态数量压缩至 10%–30% 之间, 以达到既高效又精确的状态抽象效果.

算法 1. 基于因果语义的特征区间化抽象算法.

输入: 具体状态集合 S, 语义值映射 Θ, 最大区间长度 d_{MAX}, 最小区间长度 d_{MIN}, 区间内最小具体状态数 n_{MIN}, 期望误差 e_{MEAN}, 最大误差 e_{MAX}, 缩减等级 r_d;

输出: 区间化的抽象空间 $\hat{S}_I$.

1. $\hat{S}_I \leftarrow \varnothing$ //初始化抽象过程
2. **while** !refined **do**
3. **for** each $j \in \{1,\ldots,J\}$ **do**
4. $d_j \leftarrow s$ 根据 Θ_j, d_{MAX}, d_{MIN}, n_{MIN} 进行区间化划分
5. **end for**
6. $D \leftarrow d_1,\ldots,d_J$ //形成区间化特征集合
7. $\hat{S}_I \leftarrow$ 将 D 映射为状态
8. $e_{\text{mean}}, e_{\text{max}} \leftarrow$ 计算 $\hat{S}_I$ 与 S 之间误差
9. $r_{\text{cur}} \leftarrow$ 根据$(\hat{S}_I, S)$ 计算特征压缩率
10. **if** $e_{\text{mean}} > e_{\text{MEAN}}$ or $e_{\text{max}} > e_{\text{MAX}}$ or $r_{\text{cur}} > r_d$ **then**
11. 更新 d_{MAX}, d_{MIN}, n_{MIN} //如有必要, 更新区间参数
12. **else**
13. refined $\leftarrow$ True //如果条件满足, 则结束细化
14. **end if**
15. **end while**
16. 返回 $\hat{S}_I$ //返回区间化的抽象空间

基于因果语义的特征区间化抽象不仅减少了 ICPS 中庞大的状态空间, 也为描述系统状态提供了一种人类可理解的方式, 使得设计人员在面对传感器收集的海量数据时, 依然能够直观地理解系统的特性和控制器的行为. 即便在预定义的误差参数 e_{MEAN} 设定为接近零的极端情况下, 算法 1 也能保证收敛, 即在最不利的情况下返回原始数据构建的 MDP. 通过特征区间化抽象, 具有相近语义值的具体状态被有效地映射至相同的区间化抽象状态, 为基于时空语义度量的进一步抽象提供了基础.

4.1.2 基于因果语义的时空 (ε,d)-抽象

上述基于因果语义的特征区间化抽象缩减了状态空间, 但是其简化程度仍然受到抽象粒度的影响. 为此, 我们进一步提出基于因果语义的时空 (ε,d)-抽象, 实现更加灵活和精确的模型抽象.

定义 5. 时空语义. 对于任意具体状态 $s \in S$, 时空语义 $\theta = \Theta\{V(s), Q(s,a), R(s,a), P(s,s'),\ldots\}$, 其中 $\theta \in R^n$. 这里的 θ 表示通过映射函数 $\Theta: S \to \theta$ 从状态 s 提取出的语义值, 包括了状态的多维特征, 例如状态价值函数 $V(s)$、动作价值函数 $Q(s,a)$、奖励函数 $R(s,a)$ 和迁移概率函数 $P(s,s')$ 等.

其中, 语义映射函数 Θ 用于捕获状态固有属性, 并将所处状态转化为语义空间中的坐标. 时空语义极大地丰富了对状态演变过程的认知, 提供了一个从时间和空间特征捕捉状态动态变化的分析框架. 通过这一框架, 能够更加宏观地评估不同状态之间的等价性, 实现高效抽象. 除此之外, 我们采用 (ε, d)-抽象方法[21]实现对经过因果语义区间化后的状态进一步抽象.

定义 6. (ε, d)-抽象. (ε, d)-抽象定义为一个映射:$\Phi_{\varepsilon,d}: S \to \hat{S}$, 该映射需要满足下列条件:

$$d(s_1,s_2) \leqslant \varepsilon, \forall \hat{s} \in \hat{S}, s_1,s_2 \in \Phi_{\varepsilon,d}^{-1}(\hat{s}) \tag{8}$$

其中, $\Phi: S \to \hat{S}$ 表示为抽象映射函数, 将原始状态空间 S 映射为一个抽象状态空间 $\hat{S}$. 映射函数 Φ 可以将一个真实马尔可夫模型转化为抽象模型. 令 $Pow(S)$ 表示为 S 的幂集, $\Phi^{-1}: \hat{S} \to Pow(S)$ 表示函数的逆映射. 状态抽象的核心是测量状态之间的相似性, 并根据状态相似度进行近邻抽象. 其中 d 表示状态度量矩阵, ε 表示抽象阈值.

根据马尔可夫决策过程中的状态价值函数和动作价值函数可知, 如果两个状态的迁移模型和奖励相似, 那么两个状态下的期望累积奖励也是相似的. 这为状态抽象提供了一种简化方法, 即奖励函数和迁移概率可组成该状态的时空价值矩阵, 从而在基于时空价值语义抽象的过程中, 尽可能地保持抽象马尔可夫决策过程的最优值函数,

保持与真实马尔可夫决策过程的语义等价性.

定义 7. 时空语义度量. 时空语义度量通过比较两个状态 s_1 和 s_2 在奖励、迁移概率、动作空间分布及状态上的相似性, 来量化它们之间的等价性. 对于任意 $s_1, s_2 \in S$,

$$d(s_1,s_2)=d(\theta_{s_1},\theta_{s_2}) \triangleq \max_{\hat{a}\in\hat{A}(s_1)\cap\hat{A}(s_2)} \{c_R|R(s_1,\hat{a})-R(s_2,\hat{a})|+c_P D_p[P(\cdot|s_1,\hat{a}),P(\cdot|s_2,\hat{a})]\}+c_D D_A[\hat{A}(s_1),\hat{A}(s_2)]+c_T D_S[s_1,s_2] \quad (9)$$

其中, c_R, c_P, c_D, c_T 分别为奖励差异、迁移概率差异、动作空间差异和状态差异的权重系数, 用于调节各部分对最终度量值的影响程度. $D_P[P(\cdot|s_1,a),P(\cdot|s_2,a)]$ 表示两个状态下采取相同动作后状态转移概率分布的差异度量. $D_A[\hat{A}(s_1),\hat{A}(s_2)]$ 表示两个状态可采取的动作集合的差异度量. $D_S[s_1,s_2]$ 表示状态 s_1 和 s_2 的直接差异度量. $\max_{\hat{a}\in\hat{A}(s_1)\cap\hat{A}(s_2)}$ 表示在两个抽象状态的抽象动作集合交集中能最大化后面表达式的抽象动作 $\hat{a}$. c_D, c_T 是充分大的正数, 当 $c_D D_A[\hat{A}(s_1),\hat{A}(s_2)]\leqslant\varepsilon$ 时, $\hat{A}(s_1)$ 等价于抽象动作集合 $\hat{A}(s_2)$. 时空价值矩阵满足互模拟性、唯一性, 即当 $d(s_1,s_2)=0$ 时, $s_1=s_2$.

结合定义 6 和定义 7, 利用 4 个参数来修改抽象的精度: 1) 两区间之间的最小距离 d_{MIN}; 2) 抽象状态中包含的最小具体状态数量 n_{MIN}; 3) 两区间之间的最大距离 d_{MAX}; 4) 时空价值矩阵阈值 ε.

最小距离 d_{MIN} 保证了两个不同区间反映不同的语义层次, 最小包含具体状态数量 n_{MIN} 避免了只包含少量具体状态的冗余区间出现. 此外, 当状态存在临界情况时, 将其合并到相邻的抽象状态可能会导致语义值出现显著误差, 因此, 使用第 3 个参数, 最大距离 d_{MAX}, 将临界状态分割为单独的抽象状态. 抽象状态的语义值设置为其包含的具体状态的平均语义值. 最后, 设定阈值 ε, 以第 1 阶段抽象作为其输入, 使用 (ε,d)-抽象得到最终的抽象状态空间 $\bar{S}$ 和抽象状态函数 Φ.

需要注意的是, 相关参数可以根据精度和模型尺寸要求进行修改, 即越小的误差阈值, 对应的模型越精确, 抽象状态空间越大. 公式 (7) 中第 1 个平均误差的公差 e^j_{MEAN} 是为了保证抽象的代表性和整体的准确性. 理想情况下, 每个抽象状态应该表示一组具有相似语义的具体状态, 因此每个语义的平均误差应该相对较小, 以防止抽象状态偏离其实际物理状态. 同样, 对于 e^j_{MAX}, 如果出现较大的平均误差, 则可能导致过多语义差异明显的具体状态被聚合为相同的抽象状态, 从而导致状态空间过度压缩. 第 1 阶段抽象的作用是将离散化的数据状态划分为不同的区间, 第 2 阶段的 (ε,d)-抽象可以划分某类抽象状态的语义鲁棒半径, 实现更高层级的模型抽象. 因此, 基于时空语义度量的 (ε,d)-抽象的核心目标是在最优值函数上尽可能保持抽象 MDP 与真实 MDP 的一致性, 以确保它们在语义层面上等效. 这种方法的精髓在于有效地保留对系统行为具有重大影响的关键信息, 并剔除对决策和系统性能影响不大的信息, 从而实现状态空间高效压缩和简化. 此外, 通过时空语义度量, 在空间维度上探究状态之间相似性, 并考虑了状态变化和动作选择概率分布, 在解决高维状态空间问题中提供了新视角. 这种基于语义相似性的状态度量方法使本文能够更准确地评估状态之间距离, 并深入理解系统行为本质.

算法 2 首先确定聚类数量 k (见第 1 行), 然后通过随机初始化聚类中心点 (见第 2 行) 来准备聚类过程. 接下来, 通过迭代将数据点分配给最近的聚类中心点, 并更新中心点, 直到达到收敛状态 (见第 3–10 行). 最终, 算法返回基于语义的 (ε,d)-抽象空间 $\hat{S}$ 和抽象映射函数 Φ. 这里的聚类中心点 $c_1, c_2, c_3,\ldots, c_k$ 包含了抽象空间的最终表示 (见第 11 行).

算法 2. 基于因果语义的时空 (ε,d)-抽象算法.

输入: 区间化的抽象空间 $\hat{S}_I=\{D_1,D_2,\ldots,D_n\}$, 最优状态数量确定函数 K, 时空语义度量 d;
输出: 基于语义的 (ε,d)-抽象空间 $\hat{S}$, 抽象模型 Φ.

1. 确定聚类数量 $k \leftarrow K(\hat{S}_I)$
2. 随机初始化聚类中心: $c_1,c_2,\ldots,c_k \leftarrow \hat{S}_I$ 中的随机点
3. **while** 未收敛 **do**
4. **for** i=1 to n **do**

5. 将每个数据点 D_i 分配给最近的聚类中心点: $c_{j(i)} = d(D_i, c_k)$
6. **end for**
7. **for** k=1 to k **do**
8. 更新每个聚类中心点为被分配数据点的均值: $c_k = \frac{1}{|\{i : j(i) = k\}|}\sum_{i:j(i)=k} D_i$
9. **end for**
10. **end while**
11. 返回 $\hat{S}, \Phi(c_1, c_2, \ldots, c_k)$

算法的时间复杂度取决于聚类过程的迭代次数, 其中包括将数据点分配给聚类中心点和更新中心点. 该步骤的复杂度为 $O(n \cdot k \cdot m)$, 其中 n 表示数据点数量, k 表示聚类数量, m 表示维度, 而中心点更新的复杂度为 $O(k \cdot m)$. 算法的性能受到初始中心点位置和聚类数量的影响, 因此需要根据具体数据集进行调整以获得最佳结果.

4.2 基于区间盒的动作抽象

本文基于文献 [28] 提出了一种基于区间盒 (IntervalBox) 的连续动作离散化抽象方法. 该方法的核心思想是通过对状态空间进行离散化来实现对连续动作空间的精细划分, 并利用这些抽象的动作区间来近似模拟实际策略中的动作效果. 具体为使用基于反例引导的抽象和精化 (CEGAR) 方法, 将连续的状态空间离散化为有限的抽象状态空间. 在初始阶段, 状态空间使用区间盒进行粗略的离散化, 并根据验证结果逐步对抽象状态进行细化. 通过这种方式能够有效地将原本连续的动作空间离散化, 既有助于强化学习算法处理动作选择问题, 同时也大幅降低相关计算复杂度.

定义 8. 区间盒. 对于 d 维连续动作空间 A, 每个特征中变量都有其自己的有效范围, 即第 i 个特征 ($i \in [0, \ldots, d]$) 中的变量 a_i 在范围 $[l_i, u_i]$ 内. 区间盒方法将此范围均匀划分为单位区间 $I_i = [l_i, u_i]/g_i$, 实现对连续动作空间 A 的划分. 对于动作 a, 基于区间盒抽象后, 在抽象动作空间 $\hat{A}$ 中对应动作 $\hat{a} = [k_1, k_2, \ldots, k_d]$, 其中 $k_i = a_i/g_i$, g_i 是第 i 维的抽象粒度.

动作抽象如图 5 所示, 首先均匀划分动作空间为等长的区间. k 维动作空间被划分为 m^k 个子空间, 其中每个维度上有 m 个相同长度的区间. 然后, 将具体动作转换为动作所属的区间, 即对应的抽象动作; 也就是说, 对于一个具体动作 $a \in [l,u]$, 其抽象动作为 $\bar{a}$=$[l,u]$.

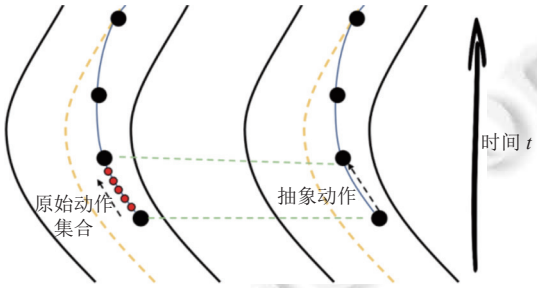


图 5 基于区间盒的动作抽象示意图

根据定义 8, 在抽象 MDP 中, 为了确保抽象动作的执行效果能够尽可能地接近真实 MDP 中相应动作的效果, 关键在于准确调整抽象粒度. 具体而言, 在确定抽象粒度时必须综合考虑环境特性和要求, 并根据具体情况设定每个状态特征的粒度级别.

更细致的粒度能够使得抽象动作更贴近原始连续动作, 从而在模拟真实 MDP 行为方面提供更高准确性. 然而, 过于微小的粒度设置也伴随着潜在风险, 可能会加剧数据中由随机波动引起的不准确性, 并影响模型的稳定性和可靠性. 因此, 在确定抽象粒度时需要找到一个适当平衡点, 以兼顾细粒度带来的近似精度和过小粒度可能导致的不稳定性. 根据具体环境特征和目标, 在不同应用场景和需求下灵活选择合适的抽象级别是必要的. 通过这种方

式既能保证抽象动作在仿真真实系统行为时高度接近, 又能避免由于错误选择粒度而导致模型不稳定和预测误差. 这一原则对于使用抽象 MDP 模型解决实际问题尤其重要, 在需要进行准确控制和决策的复杂系统中具有重要理论价值和实践意义.

4.3 构建场景的抽象 MDP

本节重点讨论在随机环境中构建该场景的马尔可夫决策过程[29], 图 6 展示了该方法的具体步骤.

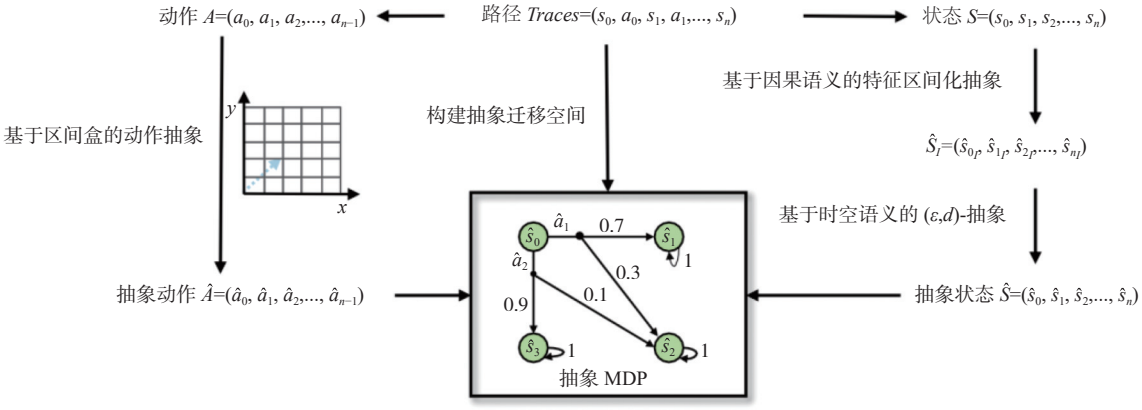


图 6 抽象 MDP 的构建过程示意图

抽象主要包括 3 个步骤: 基于因果时空语义的状态抽象、基于区间盒的动作抽象以及迁移空间构建. 通过运用算法 1 和算法 2, 状态抽象对状态空间进行压缩, 以捕捉其关键特征. 根据第 4.2 节内容, 动作抽象将真实世界中的连续多样的动作离散化为不同区间, 每个区间代表一组相似的动作. 本节将描述迁移空间的构建过程.

使用抽象状态空间 $\hat{S}$ 和抽象动作空间 $\hat{A}$, 构建抽象迁移空间 $\hat{T}:\hat{S}\times\hat{A}\to\hat{S}$, 抽象迁移空间是一组具体状态之间实际迁移的集合. 特别地, 如果存在具体状态 $s\in\hat{s}$ 和 $s'\in\hat{s}'$ 之间的实际迁移, 则相应地建立起抽象状态 $\hat{s}$ 和 $\hat{s}'$ 之间的抽象迁移, 其中 s 和 s' 是具体状态, 而 $\hat{s}$ 和 $\hat{s}'$ 是抽象状态. 抽象迁移共享相同的起始状态和目标状态, 通过迁移概率函数来表示抽象模型. 具体而言, $\eta(\hat{s},\hat{a},\hat{s}')$ 表示在当前状态 $\hat{s}$ 和当前动作 $\hat{a}$ 的条件下访问 $\hat{s}'$ 的概率, 且 $\sum_{\hat{s}'\in\hat{S}}\eta(\hat{s},\hat{a},\hat{s}')=1$. 迁移概率定义如下:

$$\eta(\hat{s},\hat{a},\hat{s}')=\frac{\left|\left\{(\hat{s},\hat{a},\hat{s}')\in\hat{T}\,|\,\hat{s}\in\hat{S},\hat{a}\in\hat{A},\hat{s}'\in\hat{S}\right\}\right|}{\left|\left\{(\hat{s},\hat{a},\_)\in\hat{T}\,|\,\hat{s}\in\hat{S},\hat{a}\in\hat{A}\right\}\right|} \tag{10}$$

换言之, 迁移概率是通过从抽象状态 $\bar{s}$ 经过执行动作 $\hat{a}$ 到抽象状态 $\hat{s}'$ 的具体迁移数量除以从抽象状态 $\bar{s}$ 出发的所有实际迁移数量来计算的.

迁移过程如算法 3 所示, 首先初始化抽象迁移空间 $\hat{T}$ 为 0, 并设置一个布尔变量 refined 为假, 表示迁移空间的构建尚未完成. 随后, 算法进入一个循环过程, 持续对迁移概率进行估计和验证, 直到满足精确度要求. 在每一次循环中, 算法遍历所有抽象状态和动作的组合 $(\hat{s},\hat{a})$, 对于每一组合, 则进一步遍历所有可能的目标抽象状态 $\hat{s}_0$, 并根据公式 (10) 计算它们之间的迁移概率 $\hat{p}$. 这一计算步骤是基于预定义的迁移事件集合 $\hat{T}$ 和当前的抽象状态与动作来执行的. 接着, 再利用霍夫丁不等式计算当前迁移概率估计误差 *error* 的同时与预设偏差阈值 ε 进行比较. 如果得到误差小于阈值 ε 的结果, 则说明当前迁移概率估计已足够准确, 并将该迁移概率及其相应状态和动作组合添加至抽象迁移空间 $\hat{T}$, 并将 refined 标记为真以指示达到了所需精确度水平. 当所有抽象状态和动作组合均经过上述检验与添加操作, 并无更多组合需要进一步优化时, 循环结束并输出所构建抽象迁移空间.

算法 3. 迁移空间构建算法.

输入: 抽象状态集合 $\hat{S}$, 抽象动作集合 $\hat{A}$, 迁移事件集合 $\hat{T}$, 偏差阈值 ε;

输出: 抽象迁移空间 $\hat{T}$.

1. 初始化抽象迁移空间 $\hat{T} \leftarrow 0$, refined $\leftarrow$ false
2. **while** !refined **do**
3. **for** $(\hat{s},\hat{a}) \in \hat{S} \times \hat{A}$ **do**
4. **for** $\hat{s}' \in \hat{S}$ **do**
5. $\hat{p} \leftarrow$ 根据公式 (8) 计算迁移概率 $(\hat{s},\hat{a},\hat{s}')$
6. **end for**
7. **end for**
8. *error* $\leftarrow$ 根据霍夫丁不等式计算
9. **for** $(\hat{s},\hat{a}) \in \hat{S} \times \hat{A}$ **do**
10. **if** *error* < 阈值 ε **then**
11. $\hat{T}.add(\hat{s} \times \hat{a} \to \hat{s},\hat{p})$
12. refined=true
13. **end if**
14. **end for**
15. **end while**
16. 返回抽象迁移空间 $\hat{T}$

5 实验分析

5.1 实验案例

本文的实验案例基于 3 个不同主题的自动驾驶场景, 涵盖了自动驾驶领域中最为常见且最重要的驾驶环境. 由于真实的实验环境限制, 我们的方法聚焦于 Carla 仿真环境, 并采用了高维数据来尽可能模拟真实环境. 通过对比分析这些实验结果, 可以更全面地理解基于时空价值语义抽象的 ICPS 在不同驾驶情境中的表现与应用.

● 车道保持辅助 (lane keeping assist, LKA). LKA[30] 是一种高级驾驶辅助系统, 旨在帮助驾驶员将车辆保持在车道内行驶. LKA 在实现自动化驾驶和提高驾驶安全性方面起着关键作用. 该系统通过测量车辆与道路中心线之间的横向偏移 d_{lat} 和相对偏航角 θ_{yaw}, 并通过调整前轮转向角 θ_{steer} 来保持车辆沿中心线行驶. LKA 的目标是使横向偏移和偏航角都趋近于 0.

● 自适应巡航控制 (adaptive cruise control, ACC). ACC[30] 是一种智能驾驶辅助系统, 其功能是自动调整车辆速度以保持与前车的安全距离. 该系统通过控制智能体的加速度 a_{ego}, 保持两车之间的相对距离 d_{rel} 始终大于安全距离 d_{safe}. 当确保安全距离后, 系统会使智能体尽可能达到用户设定的巡航速度 v_{set}. 前车的移动由前车的加速度 a_{lead} 控制, 而安全距离则基于两车的相对速度动态变化.

● 交叉路口会车辅助 (intersection crossroad assistant, ICA). ICA[30] 是一种先进的驾驶辅助系统, 旨在复杂的交叉路口环境中提供支持, 以增强驾驶安全性. ICA 系统集成了 LKA 和 ACC 的特性, 能够确定最佳的行驶速度和方向, 确保车辆在行驶过程中始终保持在正确的轨道上. 此外, ICA 还具备随机性和混合性的特性. 随机性体现在在面对复杂、难以预测的实际驾驶环境时, 系统能够做出适应并保持稳定控制. 混合性则表现在系统能够综合运用多种驾驶策略和技术, 并根据实际需要灵活切换, 以提高系统应对复杂情况的能力.

在本文中, 所设计的 ICA 场景包括两条相互垂直的双向双车道、一辆智能体车辆和多辆环境车辆, 智能体需通过左转、直行或右转等方式, 成功通过十字路口并抵达终点, 同时避免驶出道路或与环境车辆发生碰撞. 系统的观测值需涵盖 LKA 和 ACC 所包括的维度.

5.2 研究问题

为了全面评估基于因果时空语义抽象模型方法的有效性, 我们对以下两个问题进行了研究.

● 研究问题 1 (RQ1): 基于因果时空语义的抽象建模方法可否能够实现简洁性与准确性的有效平衡?

在复杂 ICPS 场景中, 尤其是状态和动作空间庞大的场景下, 构建简洁又准确的抽象 MDP 至关重要. 抽象 MDP 能够有效降低问题复杂性并提供决策依据. 然而, 抽象 MDP 往往伴随着一定程度的信息损失, 从而影响模型的准确性和决策质量. 因此, RQ1 旨在探讨基于时空价值语义构建的抽象 MDP 模型在简化环境表示与保持决策准确性之间寻求最佳平衡点的能力. 具体地, 本文将评估基于时空价值语义的抽象方法在简化 ICPS 环境复杂性的同时, 对决策过程的影响程度以及可能导致的准确性变化.

● 研究问题 2 (RQ2): 基于因果时空语义的抽象 MDP 模型在决策性能上是否能接近或达到真实 MDP 的效果? 此外, 抽象 MDP 与实际模型之间能否实现语义上的等价?

在实践中, 真实 MDP 模型会因状态空间庞大和系统结构复杂等限制, 影响模型的应用范围和效率. 因此, 如果抽象 MDP 模型能在保持与真实 MDP 语义等价性的前提下, 展现出与真实 MDP 相似乃至更优的决策性能, 则这种抽象 MDP 的实用价值和应用范围将大大提升. 因此, 研究抽象 MDP 模型是否能够达到与真实环境模型相似的决策性能, 并探讨两者之间的语义等价性, 对于评估抽象 MDP 模型的有效性和实用性具有重要意义.

5.3 实验设置

本节详细描述了实验案例的各项实验设置、实验数据收集参数和抽象模型构建参数.

在数据收集阶段, 使用基于随机网络蒸馏 (random network distillation, RND)[31] 探索的好奇心驱动强化学习方法生成 LKA、ACC 及 ICA 控制策略, 并对案例环境进行充分探索以收集系统数据. 具体而言, 在每个场景, 使用基于好奇心驱动的强化学习控制器仿真 1 000 次, 从中收集经验. 然后, 将收集到的经验以 8:2 的比例划分为建模集和验证集. 前者用于构建抽象模型, 后者用于分析抽象模型和具体模型之间的语义误差. 针对 3 个案例中涉及的深度学习网络的超参数设定如表 1 所示, 系统的 actor 学习率和 critic 学习率分别为 1.0×10^{-4}、1.0×10^{-3}, 折扣因子 γ 设定为 0.95, 策略延迟更新步长设定为 2. 此外, LKA 和 ACC 的好奇心网络损失比例为 2, ICA 的好奇心网络损失比例为 0.1. LKA 奖励设定为 $reward = 1-y^2-\theta_{\text{yaw}}^2-outOfLane_{\text{cost}}$; ACC 奖励设定为 $reward = 0.05\times v_x-collision_{\text{cost}}-outOfLane_{\text{cost}}$; ICA 奖励设定为 $reward = speed_{\text{reward}}+arrived_{\text{reward}}-collision_{\text{cost}}$. 当智能体的速度在设定范围内且在规定时间内抵达目的地时, 会给予奖励. 当智能体偏离车道中心线、与车道方向的夹角变大或发生碰撞时, 则会进行惩罚.

表 1 强化学习超参数

系统	actor学习率	critic学习率	γ折扣因子	延迟更新步长	好奇心网络损失比例
LKA	1.0×10^{-4}	1.0×10^{-3}	0.95	2	2
ACC	1.0×10^{-4}	1.0×10^{-3}	0.95	2	2
ICA	1.0×10^{-4}	1.0×10^{-3}	0.95	2	0.1

抽象过程的超参数, d_{MIN}^j 从 0.01 更新到 0.1, 步长为 0.001; d_{MAX}^j 从 0.1 更新到 0.005, 步长为 0.005; n_{MIN}^j 从 0.1% 更新到 10% 的值集合, 步长为 0.1%. 自动更新过程按照步长进行迭代, 从中选择一组最佳阈值, 更新过程中需要保证 $d_{\text{MIN}} \leqslant d_{\text{MAX}}$. 期望误差阈值 e_{pred} 设置为 0.05. 平均语义误差 e_{MEAN} 设置为 0.005, 为整个语义值域范围的 0.25%, 语义最大误差 e_{MAX} 设置为 0.2, 为语义值域范围的 10%. 优化阈值设定为 0.5%, d_p 度量概率密度的函数, 较为常见是的是康托洛维奇度量 (Kantorovich metric) 和总变异度量 (total variation metric)[32], 本文为了简化分析, 选择总变异度量衡量概率分布之间的距离.

5.4 实验结果与分析

● RQ1: 在简洁性和准确性方面, 基于时空价值语义的抽象效果如何?

为了回答 RQ1, 从压缩率 (compression ratio, *CR*) 和平均绝对误差 (mean absolute error, *MAE*) 对抽象模型进行

评估:

$$CR = \frac{|\bar{S}|}{|S|} \tag{11}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} MEAN|y-\hat{y}| \tag{12}$$

其中, $|\bar{S}|$ 表示抽象状态个数, $|S|$ 原始具体状态个数, y 是抽象模型的预测输出, $\hat{y}$ 是真实模型输出, $MEAN|y-\hat{y}|$ 测量了一次试验中偏离参考值的平均值, n 表示一次实验中抽象模型产生的抽象状态个数. CR 评价抽象模型的简洁性, 即是否具有好的抽象效果, MAE 揭示抽象模型的准确性, 即能否保留原始语义信息. (ε,d)-抽象阈值 ε 分别由 Canopy 聚类算法、肘方法 (Elbow) 和 Gap 确定.

实验结果如表 2 所示, 压缩率 (CR) 表明抽象过程能够有效地减小系统的状态、动作和转移空间. 实际上, 不同矩阵的压缩效果均达到了 99% 以上, 抽象状态模型中最多仅包含数百个不同的抽象状态, 相比之下, 真实模型中存在数万个具体状态. 对于时空价值矩阵和欧氏矩阵, 在缩小规模相近的情况下, 基于因果时空价值语义的抽象方法呈现出相对较小的平均绝对误差 (MAE), 因此可以更好地保留原始状态的价值语义信息.

表 2　不同度量矩阵对比分析

度量矩阵	ICPS系统	ε确定方法	状态个数	抽象状态个数	CR (%)	MAE
欧氏矩阵	LKA	Canopy	19 728	53	0.268 7	7.275 5
		Elbow		12	0.060 8	12.348 5
		Gap		4	0.020 7	47.625 4
	ACC	Canopy	23 756	53	0.223 1	3.362 3
		Elbow		8	0.033 6	5.893 6
		Gap		3	0.012 6	122.232 3
	ICA	Canopy	58 991	225	0.381 4	3.423 7
		Elbow		137	0.232 2	6.628 4
		Gap		26	0.044 1	48.557 5
因果时空价值矩阵	LKA	Canopy	19 728	49	0.248 3	5.177 8
		Elbow		20	0.101 4	8.677 5
		Gap		6	0.030 4	126.957 4
	ACC	Canopy	23 756	50	0.210 5	1.233 6
		Elbow		12	0.050 5	4.465 3
		Gap		17	0.071 7	3.354 5
	ICA	Canopy	58 991	302	0.511 9	1.115 8
		Elbow		156	0.264 4	3.423 6
		Gap		33	0.055 9	27.785 3

实验结果表明, 基于因果时空价值矩阵的抽象方法能够以简洁且准确的方式描述这 3 个系统的语义特征. 换言之, 基于因果时空语义的抽象模型为理解系统状态提供了一种更为简化且直观的方式.

对 RQ1 的回答: 基于时空价值语义的抽象可以有效地降低系统的复杂度, 并精确地捕获系统特征.

● RQ2: 如何保证抽象模型与真实马尔可夫模型的语义等价性?

为了回答 RQ2, 采用 PRISM 验证器建模该抽象模型, 其中, 需要保留状态的必要信息, 包括奖励、迁移概率、超出车道信息和碰撞信息等. 通过使用 PRISM 进行模型仿真, 定义与奖励和危险信息相关的性质, 以评估抽象模型在逼近真实马尔可夫模型的程度. 在此过程中, 将深入研究抽象模型的决策效果, 探究其在决策方面的优势和局限性.

表 3 中, 将不同案例的抽象模型转换为 PRISM 的验证模型, 并定义模型需要满足的性质 [33], 通过 PRISM 统计模型检测, 实现衡量基于时空价值矩阵的抽象模型与真实模型之间的语义等价性. 以十字路口场景为例, $R_{\min}$=?[C<= 60] 表示智能体在 60 步内的最小预期累计奖励, $P_{\max}$=?[F<= 60;*isOutOfLane* = 1] 表示智能体在 60 步

内至少超出一次车道的最大概率, P_{max}=?[F<= 60;$isCrashed$ = 1] 表示智能体在 60 步内至少发生一次碰撞的最大概率, 同时 P_{max}=?[F<= 60;$reachDest$ = 1] 表示智能体在 60 步内至少抵达一次终点的最大概率.

表 3 基于 PRISM 的抽象 MDP 语义误差分析

ICPS系统	抽象方法	性质	验证结果	真实值	误差
LKA	欧氏矩阵	R_{min}=?[C<= 51]	44.43	48.50	4.07
		P_{max}=?[F ⩽ 51;$isOutOfLane$ = 1]	0.13%	0.0%	0.13%
	时空价值矩阵	R_{min}=?[C<= 51]	46.92	48.50	1.58
		P_{max}=?[F⩽ 51;$isOutOfLane$ = 1]	0.10%	0.0%	0.10%
ACC	欧氏矩阵	R_{min}=?[C<= 51]	57.53	59.94	2.41
		P_{max}=?[F⩽ 51;$isOutOfLane$ = 1]	0.7%	0.0%	0.7%
		P_{max}=?[F⩽ 51;$isCrashed$ = 1]	0.06%	0.0%	0.06%
	时空价值矩阵	R_{min}=?[C<= 51]	60.33	59.94	−0.39
		P_{max}=?[F ⩽ 51;$isOutOfLane$ = 1]	0.01%	0.0%	0.01%
		P_{max}=?[F⩽ 51;$isCrashed$ = 1]	0.19%	0.0%	0.19%
ICA	欧氏矩阵	R_{min}=?[C<= 60]	8.38	9.36	0.98
		P_{max}=?[F⩽ 60;$isCrashed$ = 1]	18.73%	20.80%	2.07%
		P_{max}=?[F⩽ 60;$reachDest$ = 1]	0.17%	4.60%	4.43%
	时空价值矩阵	R_{min}=?[C<= 60]	9.38	9.36	0.02
		P_{max}=?[F⩽ 60;$isCrashed$ = 1]	19.35%	20.80%	1.45%
		P_{max}=?[F⩽ 60;$reachDest$ = 1]	2.50%	4.60%	2.09%

表 3 中的实验结果表明, 基于时空价值矩阵的方法在 R_{min} 误差 (–0.39~1.58) 和概率误差 (0.01%–2.09%) 上均优于欧氏矩阵 (R_{min} 误差为 0.98–4.07, 概率误差为 0.06%–4.43%), 更接近真实马尔可夫决策过程, 保障了抽象模型与真实模型的语义等价性.

对 RQ2 的回答: 基于时空价值语义测抽象模型更能够保证与真实模型间的语义等价性, 可以为训练策略提供有意义的指导.

6 总 结

基于深度强化学习的决策生成过程中, 其状态空间具有高维、复杂的特点, 导致其决策生成的效率低. 如何对深度强化学习的过程进行抽象建模、降低状态空间的复杂度是目前亟需解决的问题. 因此, 本文提出基于因果时空语义的抽象建模方法, 通过时间和空间变化的价值分布并结合强化学习过程中的状态价值函数, 动作价值函数及迁移概率分布等, 形成因果时空语义模型. 通过构建时空价值矩阵, 对状态进行双阶段语义抽象, 从而构建深度强化学习过程的抽象 MDP 模型, 解决现有抽象方法难以同时保留时间语义信息, 空间语义信息和概率语义信息的问题. 此外, 结合优化技术对抽象 MDP 模型进行调优, 减少抽象状态与具体状态之间的语义误差. 结合车道保持辅助系统、自适应巡航系统、交叉路口会车等案例进行实验分析, 并使用 PRISM 对模型进行验证评估, 结果表明本文所提出的抽象建模技术在模型简洁性、准确性等方面具有较好的性能优势, 同时能够保证抽象 MDP 与真实 MDP 之间的语义等价性. 但是, 由于真实驾驶环境复杂多变, 仿真与真实系统之间的差异难以避免, 本文所提方法在实际的极端复杂环境中, 仍然具有一定的局限性, 其处理高维时空数据的效率需要进一步进行实验验证.

未来的工作将继续探索基于因果时空语义构建的抽象模型在深度强化学习中的具体应用, 旨在提高基于强化学习的决策生成方法的效率和安全性. 本文所提出的抽象建模方法本质上对强化学习训练过程中, 其交互的环境进行了抽象建模. 由于物理世界的采样代价和危险性远高于仿真环境, 下一阶段, 我们将进一步验证实际环境中的抽象效果, 并充分考虑实际运行环境中的高维度和噪音问题, 引入更加细化的语义分层机制并尝试更加鲁棒的误差控制算法, 以探索基于因果时空语义抽象的强化学习在实际 ICPS 领域中的应用.

References:

[1] Radanliev P, de Roure D, van Kleek M, Santos O, Ani U. Artificial intelligence in cyber physical systems. AI & Society, 2021, 36(3): 783–796. [doi: 10.1007/s00146-020-01049-0]

[2] Li SE. Deep reinforcement learning. In: Li SE, ed. Reinforcement Learning for Sequential Decision and Optimal Control. Singapore: Springer, 2023. 365–402. [doi: 10.1007/978-981-19-7784-8_10]

[3] Junges S, Spaan MTJ. Abstraction-refinement for hierarchical probabilistic models. In: Proc. of the 34th Int'l Conf. on Computer Aided Verification. Haifa: Springer, 2022. 102–123. [doi: 10.1007/978-3-031-13185-1_6]

[4] Devidze R, Kamalaruban P, Singla A. Exploration-guided reward shaping for reinforcement learning under sparse rewards. In: Proc. of the 36th Int'l Conf. on Neural Information Processing System. New Orleans: Curran Associates Inc., 2022. 422.

[5] Kulkarni TD, Narasimhan K, Saeedi A, Tenenbaum J. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In: Proc. of the 29th Int'l Conf. on Neural Information Processing Systems. Barcelona, 2016. 3675–3683.

[6] Li LH, Walsh TJ, Littman ML. Towards a unified theory of state abstraction for MDPs. 2006. http://anytime.cs.umass.edu/aimath06/proceedings/P21.pdf

[7] Castro PS. Scalable methods for computing state similarity in deterministic Markov decision processes. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 10069–10076. [doi: 10.1609/aaai.v34i06.6564]

[8] Rafati J, Noelle D. Unsupervised subgoal discovery method for learning hierarchical representations. 2019. http://rafati.net/papers/Rafati-Noelle-2019-SPiRL.pdf

[9] Abel D. A theory of abstraction in reinforcement learning. arXiv:2203.00397, 2022.

[10] Abel D, Arumugam D, Lehnert L, Littman ML. State abstractions for lifelong reinforcement learning. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholmsmässan: PMLR, 2018. 10–19.

[11] Altman E. Constrained Markov Decision Processes. New York: Routledge, 2021. [doi: 10.1201/9781315140223]

[12] Andreas J, Klein D, Levine S. Modular multitask reinforcement learning with policy sketches. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: PMLR, 2017. 166–175.

[13] Oh J, Singh S, Lee H, Kohli P. Zero-shot task generalization with multi-task deep reinforcement learning. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: PMLR, 2017. 2661–2670.

[14] Zhang TR, Guo SQ, Tan T, Hu XL, Chen F. Generating adjacency-constrained subgoals in hierarchical reinforcement learning. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. 2020. 21579–21590.

[15] Allen C, Parikh N, Gottesman O, Konidaris G. Learning Markov state abstractions for deep reinforcement learning. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. 2021. 8229–8241.

[16] Taïga AA, Courville A, Bellemare MG. Approximate exploration through state abstraction. arXiv:1808.09819, 2018.

[17] Taylor JJ, Precup D, Panagaden P. Bounding performance loss in approximate MDP homomorphisms. In: Proc. of the 22nd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2008. 1649–1656.

[18] Feng S, Sun HW, Yan XT, Zhu HJ, Zou ZX, Shen SY, Liu HX. Dense reinforcement learning for safety validation of autonomous vehicles. Nature, 2023, 615(7953): 620–627. [doi: 10.1038/s41586-023-05732-2]

[19] Abel D, Umbanhowar N, Khetarpal K, Arumugam D, Precup D, Littman ML. Value preserving state-action abstractions. In: Proc. of the 23rd Int'l Conf. on Artificial Intelligence and Statistics. Palermo: PMLR, 2020. 1639–1650.

[20] Song JY, Xie X, Ma L. SIEGE: A semantics-guided safety enhancement framework for AI-enabled cyber-physical systems. IEEE Trans. on Software Engineering, 2023, 49(8): 4058–4080. [doi: 10.1109/TSE.2023.3282981]

[21] Guo SQ, Yan Q, Su X, Hu XL, Chen F. State-temporal compression in reinforcement learning with the reward-restricted geodesic metric. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(9): 5572–5589. [doi: 10.1109/TPAMI.2021.3069005]

[22] Bacon PL, Harb J, Precup D. The option-critic architecture. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. San Francisco: AAAI Press, 2017. 1726–1734. [doi: 10.1609/aaai.v31i1.10916]

[23] Pearl J, Mackenzie D. The Book of Why: The New Science of Cause and Effect. New York: Basic Books, 2018.

[24] Sondhi A, Shojaie A. The reduced PC-algorithm: Improved causal structure learning in large random networks. Journal of Machine Learning Research, 2019, 20(164): 1–31.

[25] Entner D, Hoyer PO. On causal discovery from time series data using FCI. In: Proc. of the 5th European Workshop on Probabilistic Graphical Models. Helsinki, 2010.

[26] Huang BW, Lu CC, Liu LQ, Hernández-Lobato JM, Glymour C, Schölkopf B, Zhang K. Action-sufficient state representation learning for control with structural constraints. In: Proc. of the 39th Int'l Conf. on Machine Learning. Baltimore: PMLR, 2022. 9260–9279.

[27] Wang ZZ, Xiao XS, Xu ZF, Zhu YK, Stone P. Causal dynamics learning for task-independent state abstraction. In: Proc. of the 39th Int'l

Conf. on Machine Learning. Baltimore: PMLR, 2022. 23151–23180.
[28] Jin P, Tian JX, Zhi DP, Wen XJ, Zhang M. Trainify: A CEGAR-driven training and verification framework for safe deep reinforcement learning. In: Proc. of the 34th Int'l Conf. on Computer Aided Verification. Cham: Springer, 2022. 193–218. [doi: 10.1007/978-3-031-13185-1_10]
[29] Hu QY, Liu JY. An Introduction to Markov Decision Processes. Xi'an: Xidian University Press, 2000 (in Chinese).
[30] Dosovitskiy A, Ros G, Codevilla F, Lopez A, Koltun V. In: Proc. of the 1st Annual Conf. on Robot Learning. PMLR, 2017. 1–16.
[31] Huang Z, Shen X, Xing J, Liu TL, Tian XM, Li HQ. Revisiting knowledge distillation: An inheritance and exploration framework. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 3579–3588. [doi: 10.1109/CVPR46437.2021.00358]
[32] Nachum O, Gu SX, Lee H, Levine S. Data-efficient hierarchical reinforcement learning. In: Proc. of the 31st Advances in Neural Information Processing Systems. Montréal, 2018. 3307–3317.
[33] Reimann J, Mansion N, Haydon J, Bray B, Chattopadhyay A, Sato S, Waga M, André É, Hasuo I, Ueda N, Yokoyama Y. Temporal logic formalisation of ISO 34502 critical scenarios: Modular construction with the RSS safety distance. In: Proc. of the 39th ACM/SIGAPP Symp. on Applied Computing. Avila: ACM, 2024. 186–195.

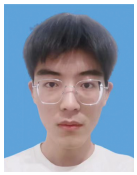
附中文参考文献:

[29] 胡奇英, 刘建庸. 马尔可夫决策过程引论. 西安: 西安电子科技大学出版社, 2000.

田丽丽(1994－), 女, 博士生, 主要研究领域为因果机器学习, 模型的可解释性.

陈逸康(2001－), 男, 硕士, 主要研究领域为机器学习, 因果推理.

杜德慧(1979－), 女, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为可信软件, 信息物理融合系统建模与验证, 人工智能安全可信理论与方法.



李荥达(2003－), 男, 本科生, 主要研究领域为强化学习, 策略生成.

聂基辉(1998－), 男, 硕士, 主要研究领域为强化学习, 形式化方法.