用于 POMCP 软策略引导的逻辑规范学习

Giulio Mazzi*维罗
纳大学 意大利维罗纳
gi
ulio.mazzi@univr.it

Daniele Meli*维罗
纳大学 意大利维罗纳
d
aniele.meli@univr.it

Alberto Castellini 维罗
纳大学 意大利维罗纳
al
berto.castellini@univr.it

Alessandro Farinelli 维罗
纳大学 意大利维罗纳
al
essandro.farinelli@univr.it

arXiv:2303.09172v1 [cs.AI] 16 Mar 2023

摘要

部分可观测蒙特卡洛规划（POMCP）是用于部分可观测马尔可夫决策过程（POMDP）的高效求解器。它采用基于蒙特卡洛树搜索的策略，在本地在线计算最优策略的近似值，从而能够扩展到大型状态空间。然而，POMCP 受稀疏奖励函数的困扰，即只有达到最终目标时才能获得奖励；在状态空间庞大、规划时域较长的环境中，这一问题尤为突出。近期，研究者将逻辑规范整合到 POMCP 中，以引导探索并满足安全要求。然而，这类与策略相关的规则需要领域专家手动定义，在现实场景中尤其如此。本文采用归纳逻辑编程，从 POMCP 执行过程的轨迹中学习逻辑规范，即规划器生成的信念-动作对集合。具体而言，我们学习以答案集编程范式表达的规则。随后，我们将这些规则整合到 POMCP 中，为有前景的动作提供软策略偏置。在两个基准场景 *rocksample* 和*battery*中，我们表明，整合从小型任务实例中学习到的规则，能够以更少的蒙特卡罗模拟提升性能，并适用于更大型的任务实例。我们公开了修改后的 POMCP 版本：
https://github.com/GiuMaz/pomcp_clingo.git。

关键词

部分可观测马尔可夫决策过程、不确定性下的规划、归纳逻辑编程、答案集编程、可解释人工智能

ACM 引用格式：
Giulio Mazzi、Daniele Meli、Alberto Castellini 和 Alessandro Farinelli。2023。用于 POMCP 软策略引导的逻辑规范学习。载于第 *22* 届自主智能体与多智能体系统国际会议（*AAMAS 2023*）论文集，伦敦，英国，*2023* 年 *5* 月 *29* 日至 *6* 月 *2* 日，IFAAMAS，共 9 页。

*作者贡献相同。

第 *22* 届自主智能体与多智能体系统国际会议论文集（*AAMAS 2023*），*A. Ricci*、*W. Yeoh*、*N. Agmon*、*B. An*（编），*2023* 年 *5* 月 *29* 日至 *6* 月 *2* 日，英国伦敦。© 2023 国际自主智能体与多智能体系统基金会（www.ifaamas.org）。保留所有权利。

1 引言

部分可观测马尔可夫决策过程（POMDP）是用于建模具有状态不确定性的系统的常用框架 [3]。智能体无法完全观测状态，因此将其建模为一种称为信念的概率分布。然而，计算最优 POMDP 策略（探索完整的信念空间）十分困难 [24]。为缓解计算方面的问题，已有多种近似算法被提出，例如部分可观测蒙特卡洛规划（POMCP） [27]和确定化稀疏部分可观测树（DESPOT） [28]。大多数现有方法仅依赖于探索信念空间的一部分，例如在 POMCP 中进行随机粒子采样；因此，必须在策略质量与探索成本之间寻求权衡。这阻碍了这些方法在信念空间庞大的现实场景中的可扩展性。此外，POMCP 策略属于黑箱，因此策略生成背后的决策过程对其他智能体和人类并不透明，降低了安全性和可信度。

本文针对上述问题，在 POMCP 的蒙特卡洛树探索中引入基于软逻辑的策略偏置。与通过设置探索过程的界限来引导探索的现有方法不同， [17, 20], 我们使用逻辑规范，根据当前信念，仅向 POMCP 求解器建议 动作。此外，我们不要求由领域专家提供这些规范。相反，我们提出从执行轨迹（即信念-动作对序列）中离线学习逻辑规范，这些轨迹可由 POMCP 或其他求解器生成。我们使用答案集编程（ASP） [18], 这一先进的逻辑规划范式 [8, 10, 21, 30]来表示规范。我们将信念转换为 ASP表示，其依据是专家用户指定的较高层次领域特征 。然后，我们利用这些特征表示信念与动作之间的逻辑关系。定义特征只需具备基础的领域知识（例如，用于表示 POMDP任务实例的相关领域量）。随后，我们利用归纳逻辑编程（ILP） [23]从特征-动作对中推断 ASP规则。

我们在两个基准部分可观测马尔可夫决策过程问题上验证了我们的方法：*rocksample* 和*battery*。本文的贡献如下：i）我们从执行轨迹中学习部分可观测马尔可夫决策过程策略的 ASP语义规范，专家只需定义领域相关特征；ii）我们使用 ASP语句引导 POMCP 探索，从而提升在具有挑战性的领域实例中的性能和可扩展性，例如规划时域增加时；iii）我们评估执行轨迹的质量和数量对

学习结果和 POMCP性能的影响。我们公开了扩展版 POMCP 算法的实现。

2 相关工作

将马尔可夫决策过程 / 部分可观测马尔可夫决策过程求解算法与逻辑推理相结合，是近年来的一种研究趋势 [17, 20, 29]。这种方法的一个优势是，可以利用常识推理和专家知识来指导策略搜索。例如，在 REBA 框架 [29] 中，答案集编程用于描述家居领域中物体与房间之间的空间关系，从而引导机器人智能体选择特定房间进行检查，同时在局部求解更简单的马尔可夫决策过程问题。类似地， [17] 的作者提出了 DARLING，它使用 ASP语句来限制模拟网格领域（类似于 rocksample）以及真实服务机器人场景中的马尔可夫决策过程探索。

逻辑语句还有助于避免不希望出现的行为，例如在安全关键场景中。为此， [19, 20] 的作者改进了规则模板，以识别意外决策，并屏蔽 POMCP 中信念-动作树的不理想分支，避免对其进行探索；研究背景是移动机器人的速度调节。 [32]，则利用命题逻辑定义一个受目标约束的信念空间，其中仅包含从初始状态安全可达的状态，并使用可满足性模理论（SMT）求解器来确保家居自动化任务得到正确执行。

上述方法以及采用时序逻辑等进行更复杂推理的类似方法 [6, 7, 16],都有一个主要缺点：它们假定逻辑语句由用户提供。这并不现实，尤其是在复杂领域中，即使是专家用户也很难定义准确的策略规则。此外，先前的研究使用逻辑语句来对马尔可夫决策过程中的额外奖励进行建模。

相比之下，我们采用答案集编程语义下的归纳逻辑编程范式，学习领域相关的用户定义特征（这些特征更容易获得）与动作之间的逻辑关系，并利用这些关系直接为 POMCP模拟提供指导。POMCP 最近已成功应用于现实场景，例如机器人规划 [11, 31, 33] 。

例如， [1, 26], 提出了学习有意义的任务表示，但使用 ASP 元程序会损害计算效率。研究表明，归纳逻辑编程有助于提升黑箱模型的可理解性 [5, 25]。机器人规划领域近期的应用包括：利用 ILASP 从人工标注示例中学习任务知识 [22],，以及从专家生成的 MDP 轨迹中学习时序规范 [6]。与例如 [26] 相比，ILASP 的效率更高，因为在最新版本中，它不依赖 ASP 元程序表示。

3 问题定义与案例研究

本文旨在通过分析执行模式，发现支配部分可观测马尔可夫决策过程智能体决策过程的逻辑规则。具体而言，智能体会根据环境的当前状态和指定任务的目标执行动作。因此，我们希望找到将环境的 POMDP 表示与特定策略相匹配的逻辑规则，以便为 POMCP 求解器提供建议，并指导智能体在未来在线执行中的行为。具体来说，通过将规则整合到

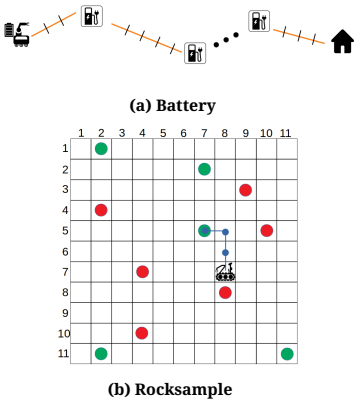


图 1：两个案例研究的示例场景。

POMCP 探索中，我们旨在提升具有挑战性的任务实例中的最终累积奖励，例如规划时域较长，或可用动作数量（以及相应所需的模拟次数）增加的情况。现在，我们将非正式介绍选作案例研究的两个领域：*battery* 和*rocksample*。在以下各节中，我们将以 rocksample 为贯穿示例，说明背景和方法的主要要素。

3.1 Battery

在电池领域中，智能体向目标位置前进，成功时获得正奖励，电池耗尽时则受到负奖励。智能体的能量自主性有限，即电池有 L 个电量级别。移动可能会降低电池电量，但路径上的已知位置设有充电站。不过，充电需要付出代价（即负奖励）。系统的可观测状态由目标、智能体和站点的位置表示，而状态的不可观测部分是电池电量，可通过噪声传感器进行检查。在此情形下，信念是对可能电池电量级别的概率分布。图 1a 展示了一个说明性场景：连接机器人、站点和目标的线段被划分为离散的前进步骤。

3.2 rocksample

在 rocksample 领域中，智能体可以在 NXN 网格上沿基本方位（北、南、东和西）每次移动一个网格单元，目标是到达并采样网格上位置已知的一组 M 岩石。岩石可能有价值，也可能没有价值。采样有价值的岩石可获得正即时奖励，而采样无价值的岩石则会带来负奖励。系统的可观测状态由智能体和岩石的当前位置描述，而在采样之前，岩石的价值是不确定的。在这种情况下，信念是关于岩石价值所有可能配置的概率分布。不过，智能体可以检查岩石的价值，检查准确度取决于岩石与智能体之间的距离。最后，智能体可以从网格右侧离开，从而获得正奖励。一个说明性场景见 $N = 11$

并 $M = 11$ 如图1b所示，其中有价值和无价值的岩石分别以绿色和红色圆点标出，智能体初始位于位置 (8，7)。

4 背景

我们首先介绍部分可观测马尔可夫决策过程和 POMCP 的定义，然后介绍答案集编程和归纳逻辑编程的关键概念，这些概念将在下文用于描述所提出的方法。

4.1 部分可观测蒙特卡洛规划

部分可观测马尔可夫决策过程（POMDP） [12]是一个元组（S、 A、 O、 T、 Z、 R、 γ），其中 S是部分可观测状态的集合， A 是动作集合， Z 是有限观测集合，T: $SXA \to \Pi(S)$ 是状态转移模型，其中 $\mathcal{B} = \Pi(S)$ 是状态上的概率分布； O: $SXA \to \Pi(Z)$ 是观测模型， R是奖励函数， $\gamma \in [0, 1]$ 是折扣因子。智能体必须最大化折扣回报 $E[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)]$。状态上的概率分布称为信念，用于表示真实状态的部分可观测性。要解决一个部分可观测马尔可夫决策过程，需要找到一个策略，即函数 π: $\mathcal{B} \to A$，将信念 $\mathcal{B}$ 映射为动作。

在这项工作中，我们重点关注部分可观测蒙特卡洛规划（*POMCP*） [27]，以求解部分可观测马尔可夫决策过程。POMCP 是一种在线算法，通过使用蒙特卡洛技术来求解部分可观测马尔可夫决策过程。POMCP 的优势在于，它不需要显式定义转移模型、观测模型和奖励，而是使用环境的黑箱模拟器。POMCP 在每个时间步使用蒙特卡洛树搜索（MCTS）探索信念空间并选择最佳动作。树置信上限（*UCT*） [13] 被用作搜索策略，以选择要探索的子树。具体而言，给定先前信念-动作对的历史 h，UCT 建议探索使动作价值最大化的动作 a：

$$V_{UCT}(ha) = V(ha) + c \cdot \sqrt{\frac{\log N(h)}{N(ha)}} \tag{1}$$

其中， V（ha）是选择动作 a 后获得的预期回报，N（h）是从历史 h 开始执行的模拟次数， N（ha）是选择动作 a 时执行的模拟次数，而 c 是探索常数。UCT 用于平衡对新动作的探索（即 N（ha）较低的动作）与对有效动作的利用（即 V（ha）较高的动作）。信念通过粒子滤波器实现，它是对可能状态的采样。信念可以随机初始化，也可以根据关于环境 [4] 的先验知识进行初始化。在每个时间步，从滤波器中选取一个粒子（代表部分可观测马尔可夫决策过程的一个特定状态），并将其用作在蒙特卡洛树中执行模拟的初始点。执行动作后，粒子滤波器会在每一步更新。一次模拟是由一系列动作—状态对构成的序列，并会累计折扣回报。通过运行不同的模拟，我们可以估算并选择能够带来最高回报的动作。如有需要，可以通过粒子复苏过程，根据当前状态生成新粒子。

4.2 答案集编程

一个 ASP 程序通过签名 和公理 [2]表示一个感兴趣的领域。签名是该领域的字母表，定义了其相关属性，包括变量（取离散值域）以及变量的谓词（原子）。例如，在 rocksample 领域中，感兴趣的变量有：岩石标识符 `R` （自然数）、智能体与岩石位置之间的距离`D` （整数），以及根据信念`V`$\in \{0, 10, ..., 100\}$[1]计算出的（离散化）概率。原子通常表示环境特征和动作。rocksample 中可能的原子有 `guess(R,V)` 和`sample(R)`，分别表示岩石`R` 有价值的概率`V`，以及采样岩石`R`这一动作。一个变量的值被赋定后，称该变量为基定 （例如，`R`=1）。如果一个原子中的变量都已基定，则该原子也是基定的。基原子会变为具有真值的布尔值。

公理是原子之间的逻辑关系。本文只考虑因果规则和弱约束。因果规则`h :-`$b_{1..n}$ 将前置条件定义为b_i 的逻辑合取（规则的规则体 ），而`h`是规则头 。例如，

```
sample(R) :- guess(R,V), V>60.
```
(2)

这意味着，如果智能体认为岩石`R` 有`V`> 60% 的概率具有价值，就可以对其进行采样。弱约束 用于表达原子之间的偏好标准。弱约束的形式如下：

$$:\sim b_1(V_1,...,V_n),...,b_m(V_1,...,V_n).[w@l,\ V_1,...,V_n]$$

其中，`w` 是权重，`l` 是（整数）优先级（当指定多个弱约束时使用），b_i 是原子，`V`j 是变量。权重可以是`V`j中的一个变量，也可以是整数。如果权重是整数，则该弱约束意味着应对b_i的基化结果施加惩罚。如果权重是变量，例如`V`$_1$,，则该约束意味着应优先选择`V`$_1$ 的较低取值。例如，如果式（2）成立，且指定了以下弱约束：

```
:~ sample(R), guess(R,V). [-V@1, R, V]
```
(3)

当有 ≥ 2 块岩石有 $> 60\%$ 的概率具有价值时，权重更高的那个（V 负权重）将更受青睐。

给定一个 ASP 任务描述，ASP 求解器会计算答案集。答案集是满足公理的最小基原子集合。从初始基化开始，求解器根据已知的基化规则体原子推导出规则的基化规则头。对于本文讨论的范围，答案集包含基化特征和动作。例如，参考式（2），如果在某个特定步骤，智能体的信念分布（以 ASP 形式化）为 {猜测(1,50), 猜测(2,70)}（即，智能体认为岩石 1 具有价值的概率为 50%，岩石 2 具有价值的概率为 70%），那么动作样本(2)将被基化，答案集将包含全部三个基原子，即 {猜测(1,50), 猜测(2,70), 样本(2)}。

4.3 归纳逻辑编程

答案集编程语义下的归纳逻辑编程问题 $\mathcal{T}$ 定义为元组 $\mathcal{T} = B, S_M, E$，其中 B 是背景知识，即答案集编程语法中的一组原子和公理（例如，变量的取值范围）； S_M 是搜索空间，即包含所有可能被学习的 ASP 公理的集合； E 是一组示例 （例如，根据执行轨迹构造的一组基原子）。我们的目标是构造一组属于 S_M 的 ASP 公理，称为假设 H，使其能够解释 E 中尽可能多的示例。为此，我们使用最先进的 ILASP 学习器 [14],，其中示例为上下文相关部分解释（CDPI）。

[1]从现在起，概率值均表示百分比，例如：{0%，10%.....，100%}。

定义 1（部分解释）。令 P 为一个 ASP 程序。P 的部分解释定义为 $e = e^{inc}, e^{exc}$，其中 e^{inc} 称为包含集，即可能属于 P 的某个答案集的基原子子集；e^{exc} 称为排除集，即不属于 P 的任何答案集的基原子子集。

定义 2（上下文相关的部分解释（CDPI））。ASP 程序 P 的 CDPI 是一个元组 e, C，其中 e 是 P 的部分解释，C 是一个称为上下文的原子集合。

本文中的部分解释包含表示动作的原子，而上下文则涉及特征原子。通过这种方式，可以学习策略规范。在我们的表述中，e^{inc} 包括已观察到的动作，而 e^{exc} 包括未观察到（因此未执行）的动作。例如，在 rocksample 中，一个 CDPI 可以是

$$\langle\langle\{\texttt{sample(2)}\},\{\texttt{sample(1)}\}\rangle,\{\texttt{guess(1,50), guess(2,70)}\}\rangle$$

其中 C ={猜测(1,50)，猜测(2,70)}，e^{exc} ={采样(1)}，e^{inc} ={采样(2)}。这个 CDPI 的含义是：给定定义智能体当前信念的上下文（也就是说，智能体认为岩石$_1$ 有价值的概率为 50%，而岩石$_1$ 有价值的概率为 70%），应该采样岩石$_2$，而不应该采样岩石$_1$。

现在，我们可以定义本文所研究的 ILASP 问题，从而明确 H 需要满足的性质：

定义 3（带 CDPI 的 ILASP 任务）。带 CDPI 的 ILASP 学习任务是一个元组 $\mathcal{T} = B, S_M, E$，其中 E 是一组 CDPI，满足：

$$\forall e = \langle\langle e^{inc}, e^{exc}\rangle, C\rangle \in E: B \cup H \cup C \models e^{inc} \wedge B \cup H \cup C \not\models e^{exc}$$

换言之，在给定环境特征上下文集的情况下，ILASP 会找出一些公理，以确保示例中观察到的 e^{inc} 中的动作能够执行（即能够在答案集中基定），而 e^{exc} 中未观察到的动作则不能执行。此外，ILASP 会找出最精简的假设 H，也就是原子数量最少的公理。这提高了已学习公理的易理解性，并提升了答案集编程求解的效率。ILASP 还可以从有序 *CDPI* 中学习弱约束，即具有预定义偏好值 [15] 的部分解释。

最后，ILASP 会找出能够解释最多示例的假设，也就是说，可能存在未被 H 覆盖的 CDPI。找到假设后，ILASP 还会返回未覆盖 CDPI 的数量，用以量化 H 相对于所提供集合 E 的可靠性。

5 方法

现在，我们将介绍如何针对给定的目标领域，从 POMDP 执行轨迹中学习策略相关逻辑规范的方法。具体而言，我们希望通过一个新的映射 $\Gamma: \mathcal{F} \to \mathcal{A}$，将策略图 $\pi: \mathcal{B} \to A$ 表示为一组逻辑公式，其中 $\mathcal{F} = \{\texttt{F}_i\}$ 是用户定义的一组分类环境特征（例如，rocksample 中的 `guess(R,V)`），而 $\mathcal{A} = \{\texttt{A}_i\}$ 是 A 的逻辑表述（例如，其中包含 rocksample 的 `sample(R)`）。假设动作和特征均表示为 ASP 原子，则构建映射 Γ 并将逻辑公式整合到 POMCP 中需要以下三个主要步骤：领域的 *ASP* 表示（第5.1节）；从 POMDP 执行轨迹中定义 ILASP 问题（第5.2节）；将已学习公理整合到 *POMCP* 中（第5.3节）。

5.1 领域的答案集编程表示

为了用答案集编程语法表示特征和动作，我们定义一个特征映射 $F_{\mathcal{F}}: \mathcal{B} \to G(\mathcal{F})$ 和一个动作映射 $F_{\mathcal{A}}: A \to G(\mathcal{A})$，其中 $G(\cdot)$ 是基化函数，用于定义一个原子或一组原子的所有可能基化结果。例如，在有2块岩石的 rocksample 场景中，考虑 $\mathcal{A}$ ={`sample(R)`}，则 $G(\mathcal{A})$ ={`sample(1)`, `sample(2)`}。用户定义映射 $F_{\mathcal{F}}, F_{\mathcal{A}}$ 后，就可以自动将部分可观测马尔可夫决策过程的执行轨迹（即信念-动作对序列）转换为基原子集合。

5.2 根据部分可观测马尔可夫决策过程轨迹定义 ILASP 问题

现在，我们可以按照第4.3节定义一个 ILASP 任务，其中的示例由部分可观测马尔可夫决策过程轨迹通过映射 $F_{\mathcal{F}}, F_{\mathcal{A}}$ 构建，目标假设为 $H = \Gamma$。为定义 ILASP学习问题，我们作出以下假设：

- 示例取自“良好”的执行轨迹，即策略近乎最优、智能体成功完成任务并获得较高折扣回报的轨迹。这样一来，学习得到的规则在在线执行时效率更高。第6节将详细说明如何生成用于学习的部分可观测马尔可夫决策过程轨迹，以满足这一假设（例如，调整 POMCP 参数）；
- 对于每个动作，相应的映射 Γ 不依赖于其他动作的映射，即每个动作的公理彼此独立（例如，用于检查岩石或采样岩石的规则相互独立，但都与特征相关联）。因此，我们可以为每个动作分别定义一个 ILASP 任务；由于搜索空间缩小，计算效率也随之提高。

每个 ILASP 问题由背景知识 B、搜索空间 S_M 和示例 E 定义。在本文中，B 仅包含 ASP 变量和取值范围的定义。由于 $\Gamma: \mathcal{F} \to \mathcal{A}$，我们定义 S_M，仅允许候选公理的规则头中出现原子 $\in \mathcal{A}$，并允许规则体中出现原子 $\in \mathcal{F}$。示例是从使用映射 $F_{\mathcal{F}}$ 和 $F_{\mathcal{A}}$ 的部分可观测马尔可夫决策过程执行轨迹中提取的。具体而言，对于每个信念—动作对 $b, a \in \mathcal{B}XA$，我们得到一个由基特征集和动作组成的对 信念，动作，其中 信念$\subseteq G(\mathcal{F})$，且 动作$\subseteq G(\texttt{A}_i)$，$\texttt{A}_i$ 是表示 a 的 ASP 原子。例如，在一个包含 2 块岩石的岩石采样实例中，假设对第二块岩石进行采样，且岩石 1 和 2 有价值的概率分别为 70% 和 80%。那么，$\texttt{A}_i$ =`sample(R)`，动作=`sample(2)` 以及 信念={`guess(1,70)`, `guess(2,80)`}。随后，我们生成形式如下的 CDPI

$$\langle\langle\{\texttt{action}\}, G(\texttt{A}_i) \setminus \{\texttt{action}\}\rangle, \texttt{belief}\rangle \tag{4}$$

其中 $G(\texttt{A}_i) \setminus \{$动作$\}$ ={样本(1)} 如前一个示例所示。此外，在每个未执行动作 a 的步骤中，我们定义一个额外的 CDPI

$$\langle\langle\emptyset, G(\texttt{A}_i)\rangle, \texttt{belief}\rangle \tag{5}$$

例如，在前一个示例中，如果 a 是检查动作，$G(\mathtt{A}_i)$ ={检查(1), 检查(2)}，其他动作也类似。这样，学习得到的公理将提供更有意义且实用的策略规范，因为它们也会从反例（即未执行动作的示例）中归纳得出。

为了学习弱约束（即相对于其他动作而言，对特定动作的偏好），对于每个如（4）所示的示例，我们生成如下形式的 CDPI:

$$\langle\langle G(\mathtt{A}_i) \setminus \{\mathtt{action}\}, \emptyset\rangle, \mathtt{belief}\rangle \tag{6}$$

换言之，所有未执行的动作都属于 e^{inc}，而 e^{exc} =∅。随后，我们指定（4）优先于（6）。这是必要的，因为执行轨迹在每个步骤只包含一个动作，但实际上可能还有其他动作可以执行。例如，考虑前述岩石采样实例，并假设（2）成立。对两块岩石进行采样都是可行的，因为它们具有 ≥ 60% 的概率有价值，但实际采样的只有第二块岩石。然后，我们定义以下示例，分别对应（4）至（6）：

$$\langle\langle\{\mathtt{sample(2)}\}, \{\mathtt{sample(1)}\}\rangle, \{\mathtt{guess(1,70),\ guess(2,80)}\}\rangle \tag{7a}$$

$$\langle\langle\{\mathtt{sample(1)}\}, \emptyset\rangle, \{\mathtt{guess(1,70),\ guess(2,80)}\}\rangle \tag{7b}$$

然后，我们指定（7a）必须优先于（7b）。

对于每个动作，我们考虑一个能够覆盖大多数示例的假设，并用它来引导 POMCP 探索，具体说明见下一小节。

5.3 在 POMCP 中整合公理

我们将计算得到的公理整合到 POMCP 中，以帮助算法根据当前状态和信念选择最佳动作。具体而言，在蒙特卡洛树搜索（MCTS）过程中到达一个新节点时，我们会用根据当前信念计算出的值来实例化公理中的变量。POMCP 使用粒子滤波器表示根节点的信念（即，算法为每个信念保留有限数量的可能状态）。因此，我们通过分析所有粒子来计算基化 ASP 特征所需的值。例如，要计算岩石$\mathtt{guess(R,V)}_1$，我们会计算岩石$_1$ 有价值的粒子所占的比例。当到达新的非根节点时，没有可用的粒子滤波器。不过，我们仍会对仅依赖于状态可观测部分的特征进行基化（例如，在 rocksample 中的当前位置）。这些特征只需在先前访问过、后来成为根节点的非根节点上重新基化（即，我们通过选择一个动作并接收一次观测，将模拟向前推进一步，而新的根节点是此前在模拟过程中访问过的节点）。在这种情况下，会使用新的完整信念重新实例化这些特征。其他情况下则无需重新计算，因为节点的可观测部分始终不变。节点的所有特征实例化完成后，我们会调用答案集编程求解器来基定新的动作。

为了将逻辑规则与 POMCP 结合，我们在 UCT 中引入一个先验（公式 1）。具体而言，对于 ASP 求解器基定的每个动作 a，我们将 $V(ha)$ 的值提高到与 c 相同的值，并将 $N(ha)$ 提高到一个表示高奖励的固定值（这取决于领域；在我们的领域中，我们根据经验将其设为 10）。这相当于假设我们已经在当前信念下进行了 $N(ha)$ 次模拟，并且取得了较好的回报值（即，$V(ha)$ 的值较高）。需要注意的是，这个先验并不会影响策略的最优性。事实上，$N(ha)$ 的数量是有限的，因此，与标准 POMCP 一样，进行无限次模拟后，我们将收敛到最优策略。然而，这个先验显著降低了蒙特卡洛树搜索期间对探索的需求，因为在蒙特卡洛树搜索开始模拟之前，公理就会指出最有可能的最佳动作。随着树规模增大（因此需要进行许多模拟），这一优势尤为关键，如第6节所示。

6 实证评估

现在，我们介绍 rocksample 和 battery 这两个案例研究的实证评估。

6.1 实证方法

我们在每个领域的 1000 个随机场景中运行 POMCP，生成学习执行轨迹，具体设置如下：$12X12$对于 rocksample，使用含有 4 块岩石的网格，并随机设置岩石的价值和位置以及智能体的初始位置；对于 battery，使用 35 步的路径，并随机设置站点位置，同时限制站点之间的距离范围（$\in [1, 4]$）。我们在 POMCP算法中设置 2_{15} 个粒子[2] 和模拟次数，以生成“良好”的执行轨迹，具体说明见第5.2节。ILASP 示例仅从折扣回报不低于所有轨迹平均值的执行轨迹中提取。这一标准与领域无关，可用于筛选出有意义的学习示例。

我们在两种情形下评估 POMCP 有无 ASP公理偏置时的性能：i）减少粒子数量，同时保持问题维度不变（针对每个粒子数量生成 50 个随机设置）；ii）固定粒子数量为 2_{15}，同时增加问题维度。具体而言，在第二种情形中，我们增大 rocksample 的网格大小N 以及 battery 的路径长度，并针对每个参数值运行 50 次随机模拟。由此，我们展示了规则在不同领域实例中的可扩展性及其实用价值。在实际应用中，规划时域对 POMCP 探索至关重要。ASP 求解使用先进的 Clingo [9]完成。所有实验均在配备 Intel Core i7-6700HQ 和 16 GB 随机存取存储器的笔记本电脑上运行。

6.2 电池

以下是我们的方法（第 5 节）在电池领域各步骤中的结果。

6.2.1 ASP 表示。我们为电池领域定义以下特征：

- $\mathtt{guess(L,V)}$，表示电池电量至少为$\mathtt{L}\in$时的概率$\mathtt{V}\in\{0, 10, ..., 100\}$%；其中$\mathtt{L}\in\in\{0, 1, ..., 10\}$。
- $\mathtt{dist\_next(D)}$，表示到下一个站点（若有）或目标的距离$\mathtt{D}$（整数）。• $\mathtt{X} \geq \bar{x}$和$\mathtt{X} \leq \bar{x}$，其中$\mathtt{X}$ 可以是$\mathtt{L,D,V}$之一，而 $\bar{x}$ 是$\mathtt{X}$的一个可能取值。
- $\mathtt{at\_station}$，表示智能体位于站点位置。

动作的原子是充电、检查、前进。

[2]按照标准做法，粒子数量与模拟次数相同。

6.2.2 ILASP 结果。ILASP 从一组 ≈ 29600[3] 示例中学习到以下因果规则（即第4.3节中的 CDPI）：

检查 :- V ≥ 20; L ≤ 4; 猜测(L, V)。 前进 :- 距离_下一个(D); D ≤ 4。 充电 :- L ≤ 7; V ≥ 30; D ≥ 2; 距离_下一个(D); 猜测(L, V); 在_站点。

当电池电量明显偏低（≤ 4）且概率不可忽略（≥ 20%）时，检查是合适的。当电池电量为 ≤ 7 且概率较高（≥ 30%）时，应进行充电，并且下一个站点不够近（D≥ 2）。最后，系统始终建议前进，因为根据我们的问题定义，下一个站点（或目标）始终在 4 步以内。最长的充电 规则（所需时间较长）的学习时间为 ≈45秒，其中 S_M 包含 ≈ 4700 公理[4]。

6.2.3 POMCP 集成结果。在图2a中，我们展示了当粒子数在 2_8 至 2_{15}之间变化时，POMCP（采用和不采用规则偏置）所取得的折扣回报。引入学习得到的 ASP 公理并不会显著影响求解器的性能。这可能是因为任务涉及的动作数量有限（仅 3 个，即前进、检查 和充电），因此每个模拟步骤只需探索少数分支。

图2b展示了当路径长度增加时，POMCP所获得的折扣回报；粒子数始终保持为2_{15}，与训练设置相同。在这种情况下，采用学习到的规则后，求解器的表现显著更好（当路径长度为75时，折扣回报最高增加20%，几乎是训练场景的两倍）。事实上，路径越长，电池耗尽的风险越高，因为未能充电的可能性也更大。此外，规划时域随之延长，模拟深度也会增加，因此ASP公理的引导至关重要。

6.3 Rocksample

下面介绍我们的方法（第5节）在rocksample 领域各个步骤中的结果。

*6.3.1 ASP*表示。 我们定义以下特征集：

- guess(R,V)，表示岩石R 为“优质”的概率V。由于答案集编程只能处理整数变量，因此V 在{0，10，.....，100}范围内进行离散化，含义如下：guess(1, 10) 表示岩石1 有价值的概率为 $P(1) \in [10, 20[$ %。我们根据经验选择了这一概率离散化步长；也可以采用更精细的离散化，但在大型领域中可能会影响答案集编程的计算时间。
- dist(R,D) 表示智能体与岩石R之间的 1-范数距离D。
- min_dist(R) 表示岩石R 距离智能体最近。
- delta_x(R,D) （分别为，delta_y(R,D)），表示岩石R 相对于智能体的D坐标（分别为D坐标）。
- X ≥ $\bar{x}$ 和X ≤ $\bar{x}$，其中X 可以是V,D，而 $\bar{x}$ 是X的一个可能取值。
- sampled(R) 表示岩石R 已采样。
- num_sampled(N) 表示已有百分之N 的岩石被采样（N ∈{0, 25，.....，100}。
- target(R) 表示岩石R 是下一个要采样的目标。

具体而言，min_dist 和best_guess 是定性特征，加入它们可能有助于简化答案集编程规则表达式并提高其可理解性。

动作原子包括east、west、north、south、exit[5]、check(R)、sample(R)，其中参数R 表示要检查或采样的目标岩石。

6.3.2 ILASP 结果。ILASP 从一组 ≈ 8500示例中学习到以下公理：

东 :- target(R), delta_x(R,D), D ≥ 1. 西 :- target(R), delta_x(R,D), D ≤-1. 北 :- target(R), delta_y(R,D), D ≥ 1. 南 :- target(R), delta_y(R,D), D ≤-1. (8d)target(R) :- 距离(R,D), not 已采样(R), D ≤ 1. 检查(R) :- target(R), not 已采样(R), 检查(R) :- 猜测(R,V), not 已采样(R), 样本(R) :- target(R), 距离(R,D), D ≤ 0, 退出 :- 猜测(R,V), V ≤ 40, not 已采样(R), 退出 :- 距离(R,D), 5 ≤D ≤ 8, not 已采样(R),

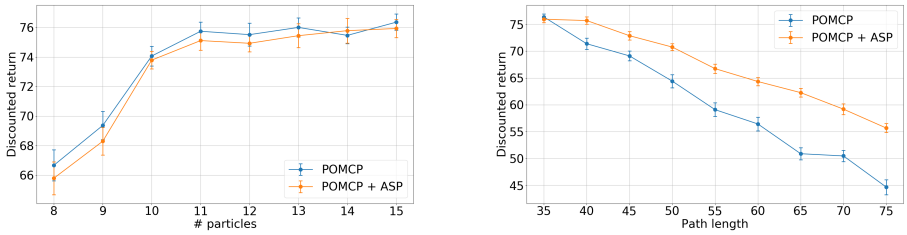
(8a) (8b) (8c) (8e)

```
target(R) :- guess(R,V), not sampled(R),
             70 ≤ V ≤ 80.                    (8f)

             guess(R,V), V ≤ 50.             (8g)

             dist(R,D), D ≤ 0, V ≤ 80.       (8h)

             not sampled(R), guess(R,V),
             V ≥ 90.                         (8i)

          num_sampled(N), N ≥ 25.            (8j)

          num_sampled(N), N ≥ 25.            (8k)
```

移动动作的公理相当简单。运动取决于目标岩石及其相对位置（delta_x 和 delta_y 原子）。关于target，发现了两条公理：若一块尚未采样的岩石与智能体之间的距离较短（≤ 1），或者它很有价值的概率较高（猜测值 V 在 70% 到 80% 之间），则将其选为目标。随后，如果智能体位于目标岩石的位置，且该岩石有 ≥ 90% 的概率很有价值，就会对其进行采样。关于check也发现了两条公理：当目标岩石很有价值的概率较低（≤ 50%）时会进行检查，或者智能体在岩石所在位置时会检查以验证其价值。最后，当至少有 25% 的岩石已采样，并且一块尚未采样的岩石很有价值的概率较低（≤ 40%）或离智能体较远（距离在 5 到 8 之间）时，智能体决定退出网格。

[3]示例的数量取决于执行中的步骤数。[4]其大小取决于特征、变量及其可能取值的数量。

[5]退出 为一般性起见，被视为一个独立的动作。由于在我们的领域中，智能体从网格右侧退出，因此在本文讨论范围内，我们手动将其映射为 POMDP 表示中的东 动作。



(a) (b)

图2：电池领域（均值 ± 标准差）。a. 减少 POMCP 粒子数量的测试；b. 路径长度逐渐增加时的测试。

ILASP 还能够为`target` 原子学习以下一组弱约束：

```
: ~ target(R), dist(R,D).[D@1, R, D]
: ~ target(R), min_dist(R), guess(R,V).
    [-V@2,R,V]
```

换言之，当根据因果规则有多个岩石可作为目标时，智能体会对它们进行排序，首先考虑距离远近，其次考虑它们具有价值的概率。对于最长的目标规则（带有弱约束），学习时间为 ≈420秒，其中包含 ≈ 29500 条公理，计有S_M。

6.3.3 POMCP 集成结果。图3a展示了当粒子数量在 2_8 到 2_{15} 之间变化时，POMCP（带规则偏置和不带规则偏置）的性能；网格大小和岩石数量与训练设置相同（$N = 12$ 和 $M = 4$）。使用较少粒子（即 $\leq 2^{11}$）时，使用规则和不使用规则所获得的折扣回报非常相似。而粒子数量增加时（即 $> 2^{11}$），学习到的规则可提升性能（使用 2_{15} 个粒子时最多提升 $15 - 20\%$）[6]。事实上，使用更多粒子时，规则对策略图的描述更加准确，因为信念分布的精度更高，因此关于岩石距离及其具有价值的概率的上下界也更加精确（例如，（8i）中的`D≤0` 和`V≥90`）。由于 rocksample 中的动作数量多于电池领域，已学习公理显著减少了 POMCP 探索中蒙特卡洛树需要扩展的分支数量。这一点在图3b中更加明显：网格大小从12增至24，而岩石数量和模拟次数保持与训练集相同，即分别为 $M = 4$ 和 2_{15}。在这种情况下，规划时域变长，规则对于取得更好的性能至关重要。事实上，在 $24X24$ 网格中，带规则的 POMCP 平均获得的折扣回报几乎是标准 POMCP 的两倍。

6.4 进一步考量

现在，我们分析流水线的几个关键方面，具体而言，是答案集求解对 POMCP 的计算影响（第 6.4.1 节），以及训练集中执行示例的数量和质量如何影响学习和规划结果（第 6.4.2–6.4.3 节）。为简洁起见，我们只考虑 rocksample 场景，事实证明，该场景比 battery 更具挑战性，因为它包含更多动作。

6.4.1 答案集编程在 *POMCP* 中的计算影响。 如第5.3节所述，蒙特卡洛树搜索中的每个节点都会进行答案集编程求解。在一个包含 2_{15} 个粒子、网格上有 4 块岩石的 rocksample 场景中，Clingo 在 50 个任务实例中每次试验平均被调用 125000 次。每次答案集编程求解调用约需 0.3 毫秒。与相同场景下纯 POMCP 求解器相比，这会增加每个步骤（即已执行的动作）的计算时间（平均 $5.53s$ 对比 $1.88s$），也会增加所有任务实例的总计算时间（平均 $27s$ 对比 $18s$）。使用 Clingo 的优点是能够保持实现的模块化，因此可以轻松地在单独的 ASP 文件中修改任务公理，并在运行时进行评估。然而，直接将 ASP公理作为条件语句写入 POMCP 代码中（示例见相关代码库）会更高效，计算时间与纯 POMCP 相近（在此前的条件下，每个任务实例平均 $20s$ 对比 $18s$）。此外，图3a 显示，与纯 POMCP 相比，公理使 $\geq 2^{12}$ 粒子能够获得更高的折扣回报，纯 POMCP 使用 $N = 2^{15}$。POMCP+Clingo 使用 2_{12} 个粒子时，平均在 $\approx 2.5s$ 内完成一个任务实例，比使用 2_{15} 个粒子的纯 POMCP 快近一个数量级。因此，我们的流水线能够以更低的计算成本获得更好的性能。

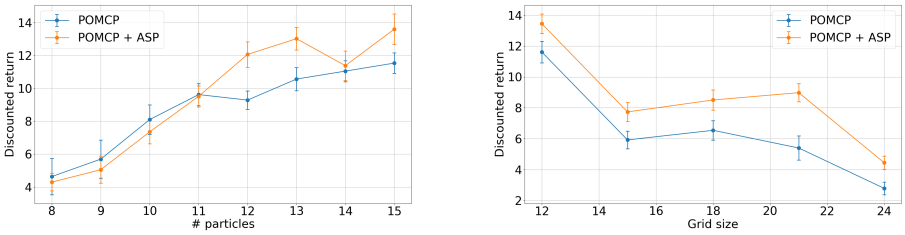
6.4.2 已学习公理的鲁棒性。 现在，我们评估训练集中示例数量对 ILASP 学习到的公理的影响。图4a–4b展示了规则学习结果如何取决于所用示例（即 CDPI）的数量。在 x 轴上，我们给出相对于先前实验所用原始训练集的示例百分比（即 ≈ 8500 rocksample 中的比例）。在图4a的 y 轴上，我们展示未覆盖示例的百分比；在图4b的 y轴上，我们展示基于给定数量的示例学习到的公理与基于完整训练集学习到的公理之间的距离（按后者中的原子数量归一化）。给定两条规则 R_1，R_2，每条规则均由一组原子 {a

i $i \in$ 1, 2 $R_1 - R_2$ =} { }

| 1 ∪ 2 | − | 1 ∩ 2 | | |

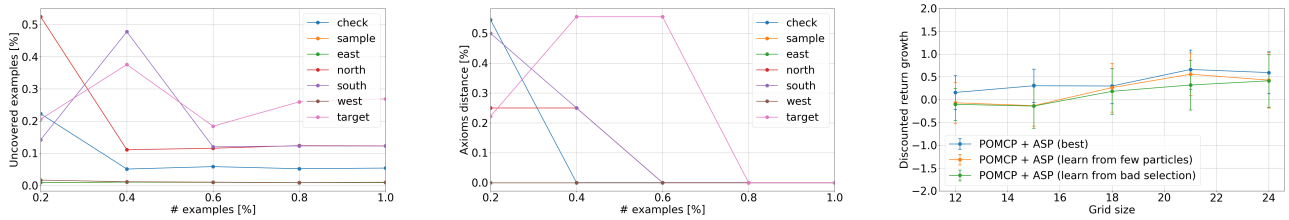
组成，我们将它们之间的距离定义为 {a} {a} {a} {a}，其中 · 表示集合的基数。例如，`sample(R) :- dist(R,V), V≤2` 与 (8i) 的距离为 5，这是因为缺少`not sampled(R)`、`guess(R, V),V ≥90` 和 `target(R)`，且距离上界`D`不同。我们观察到，使用完整训练数据集的 $\geq 80\%$ 时，未覆盖示例的百分比趋于稳定（图4a），且所有动作的距离都变为零（图4b），因此学习到的规则成功

[6] 2^{14} 粒子处的下降，是因为随机场景中有价值的岩石平均数量较少。



(a) (b)

图3：Rocksample 领域（均值 ± 标准差）。a. 使用较少 POMCP 粒子的测试；b. 网格大小增大时的测试N。



(a) (b) (c)

图4：a. 使用不同数量（#）示例进行学习时的未覆盖示例（%）；b. 使用不同数量（#）示例学习得到的规则与使用完整训练集学习得到的规则之间的相对距离；c. 网格大小较大的 rocksample 中，基于最佳示例（2_{15} 个粒子且折扣回报 >平均值）或次优示例（折扣回报 < 平均值，或由 POMCP 使用2_{11} 个粒子生成）学习公理时，折扣回报的增长率（均值 ± 标准差）。

收敛到稳定的假设。总体而言，基于完整数据集学习到的规则覆盖的示例多出 $> 73\%$。

6.4.3 示例质量对 *POMCP* 性能的影响。现在，我们评估示例质量（即生成示例的策略）对结合 ASP 公理的 POMCP 所实现的折扣回报的影响。我们想展示这样一种情况：由于任务复杂，POMCP 无法为训练集生成计算出最优策略，因此只能使用次优策略。测试时，我们首先使用 2_{15} 个粒子生成的 POMCP 最优执行轨迹来学习 rocksample 的规则，但只选择折扣回报低于所有轨迹平均值的轨迹。这模拟了难以轻易定义合适的示例选择标准的情况，例如在复杂领域中。随后，在另一项测试中，我们使用 2_{11} 个粒子生成的轨迹来学习规则，也就是说，在这种情况下，公理能够提升纯 POMCP 的性能；我们选择折扣回报为平均值 $>$ 的轨迹。这也有助于在复杂领域中快速生成示例。图4c展示了两项测试的结果（针对每个网格大小值考虑 50 种随机设置），指标为折扣回报的增长率，即使用规则与不使用规则时所实现折扣回报之间的相对差异，并以不使用规则时所实现的折扣回报进行归一化。在图表中，为了验证，将 POMCP 与规则结合运行时使用 2_{15}个粒子。总体而言，从次优示例中学到的规则不会降低纯 POMCP 的性能（网格大小为 $N < 18$ 时，平均值仅略微下降 $< 14\%$），这得益于软引导方法。

有趣的是，在两项测试中学到的次优 ASP 规则仍然有助于 POMCP，且随着规划时域增加，这种效果依然存在（$N \geq 18$）。具体而言，在 $N = 24$ 时，折扣回报提高约 50%。

7 结论

我们提出了一种用于 POMCP 软策略引导的新方法，该方法基于归纳逻辑编程，直接从部分可观测马尔可夫决策过程执行轨迹中学习描述策略规范的答案集编程规则。我们的规则无需专家手工编写；我们只需定义相关的高层领域专用特征，以便用答案集编程形式体系描述信念信息，并与部分可观测马尔可夫决策过程动作进行匹配。我们证明，在 rocksample 和 battery 基准领域中，学习得到的答案集编程公理能够显著支持 POMCP 探索，在 POMCP 模拟中引导系统探索良好的分支，同时不会显著增加计算负担。特别是，当可能动作的数量增加（rocksample）以及规划时域延长时，规则更有益。此外，即使公理是从低质量策略生成的轨迹中学习得到的，POMCP 的渐近最优性仍得以保留；并且随着示例数量增加，公理会收敛至不动点。

未来，我们计划研究：i）如何整合更先进的（例如时序）逻辑规范，以应对更复杂的任务；ii）如何处理训练

集中的未覆盖示例，并在执行过程中可能对公理进行改进；iii）用户对环境特征定义不当所产生的影响。

致谢

本项目获得意大利大学与研究部的资助，资助项目为 PON “Ricerca e Innovazione” 2014-2020”（资助协议编号：40-G-14702-1）。

参考文献

[1] Blai Bonet 和 Hector Geffner。2020。根据状态空间结构学习用于规划的一阶符号表示。在 *ECAI 2020*。IOS出版社，2322–2329。[2] Francesco Calimeri、Wolfgang Faber、Martin Gebser、Giovambattista Ianni、Roland Kaminski、Thomas Krennwallner、Nicola Leone、Marco Maratea、Francesco Ricca 和 Torsten Schaub。2020。ASP-Core-2 输入语言格式。《逻辑编程理论与实践》20, 2 (2020), 294–309。[3] Anthony R Cassandra、Michael L Littman 和 Nevin Lianwen Zhang。2013。增量剪枝：一种针对部分可观测马尔可夫决策过程的简单、快速且精确的方法。*arXiv* 预印本 *arXiv:1302.1525* (2013)。[4] Alberto Castellini、Georgios Chalkiadakis 和 Alessandro Farinelli。2019。状态变量约束对部分可观测蒙特卡洛规划的影响。在 *IJCAI 2019*, 中国澳门，*2019* 年*8*月*10–16*日。ijcai.org，5540–5546。[5] Fabio A D’Asaro、Matteo Spezialetti、Luca Raggioli 和 Silvia Rossi。2020。迈向一种用于解释黑箱偏好学习系统的归纳逻辑编程方法。在 知识表示与推理原理国际会议论文集，第17卷。855–859。[6] Giuseppe De Giacomo、Marco Favorito、Luca Iocchi 和 Fabio Patrizi。2020。利用约束螺栓对异构智能体进行模仿学习。在 自动规划与调度国际会议论文集，第30卷。517–521。[7] Giuseppe De Giacomo、Luca Iocchi、Marco Favorito 和 Fabio Patrizi。2019。约束螺栓基础：使用 LTLf/LDLf 约束规范进行强化学习。在 自动规划与调度国际会议论文集，第29卷。128–136。[8] Esra Erdem 和 Volkan Patoglu。2018。答案集编程在机器人学中的应用。*KI*-人工智能32, 2 (2018), 143–149。[9] Martin Gebser、Roland Kaminski、Benjamin Kaufmann、Max Ostrowski、Torsten Schaub 和 Philipp Wanko。2016。借助 clingo 5 轻松实现理论求解。在 第*32*届逻辑编程国际会议（*ICLP 2016*）技术交流。达格施图尔宫莱布尼茨信息学中心。[10] Michele Ginesi、Daniele Meli、Andrea Roberti、Nicola Sansonetto 和 Paolo Fiorini。2020。机器人手术中的自主任务规划与态势感知。在 *2020 IEEE/RSJ*智能机器人与系统国际会议（*IROS*）。电气电子工程师学会，3144–3150。[11] Francesco Giuliari、Alberto Castellini、Riccardo Berra、Alessio Del Bue、Alessandro Farinelli、Marco Cristani、Francesco Setti 和 Yiming Wang。2021。POMP++：基于POMCP的主动视觉搜索，适用于未知室内环境。在*IEEE/RSJ*智能机器人与系统国际会议（*IROS*）（捷克共和国布拉格）。IEEE出版社，1523–1530。[12] Leslie Pack Kaelbling、Michael L. Littman 和 Anthony R. Cassandra。1998。部分可观测随机领域中的规划与行动。人工智能101, 1 2 1998 99 134 –（5月），–。[13] Levente Kocsis 和 Csaba Szepesvári。2006。基于老虎机的蒙特卡洛规划。在 *ECML’06* 论文集。施普林格出版社，柏林，海德堡，282–293。[14] Mark Law、Alessandra Russo 和 Krysia Broda。2015。用于学习答案集程序的 ILASP 系统。www.ilasp.com。

[15] Mark Law、Alessandra Russo 和 Krysia Broda。2016。根据上下文相关示例迭代学习答案集程序。《逻辑编程理论与实践》16, 5-6 (2016), 834–848。[16] Matteo Leonetti、Luca Iocchi 和 Fabio Patrizi。2012。有限状态控制器的自动生成与学习。在 人工智能国际会议：方法、系统与应用。施普林格，135–144。[17] Matteo Leonetti、Luca Iocchi 和 Peter Stone。2016。融合自动规划与强化学习，实现高效、稳健的决策。人工智能241 (2016), 103–130。[18] Vladimir Lifschitz。1999。答案集规划。在 逻辑编程与非单调推理国际会议。施普林格，373–374。[19] Giulio Mazzi、Alberto Castellini 和 Alessandro Farinelli。2021。在部分可观测蒙特卡洛规划中识别意外决策：一种基于规则的方法。在 第*20*届自主智能体与多智能体系统国际会议论文集（*AAMAS*）。国际自主智能体与多智能体系统基金会，889–897。[20] Giulio Mazzi、Alberto Castellini 和 Alessandro Farinelli。2021。用于部分可观测蒙特卡洛规划的基于规则的屏蔽。在 自动规划与调度国际会议论文集（*ICAPS*），第31卷。243–251。[21] Daniele Meli、Hirenkumar Nakawala 和 Paolo Fiorini。2023。用于机器人审议式任务规划的逻辑编程。人工智能评论(2023), 1–39。[22] Daniele Meli、Mohan Sridharan 和 Paolo Fiorini。2021。为自主手术任务规划归纳学习答案集程序。机器学习110, 7 (2021), 1739–1763。[23] Stephen Muggleton。1991。归纳逻辑编程。新一代计算8, 4 (1991), 295–318。[24] Christos H Papadimitriou 和 John N Tsitsiklis。1987。马尔可夫决策过程的复杂性。运筹学数学12, 3 (1987), 441–450。[25] Johannes Rabold、Michael Siebers 和 Ute Schmid。2018。利用归纳逻辑编程解释黑盒分类器——借助 Aleph 增强 LIME，以关系规则近似非线性决策。在 归纳逻辑编程国际会议。施普林格，105–117。[26] Ivan D Rodriguez、Blai Bonet、Javier Romero 和 Hector Geffner。2021。从黑盒状态学习用于规划的一阶表示：新成果。在 知识表示与推理原理国际会议论文集，第18卷。539–548。[27] David Silver 和 Joel Veness。2010。大型部分可观测马尔可夫决策过程中的蒙特卡洛规划。神经信息处理系统进展23 (2010)。[28] Adhiraj Somani、Nan Ye、David Hsu 和 Wee Sun Lee。2013。DESPOT：采用正则化的在线部分可观测马尔可夫决策过程规划。神经信息处理系统进展26 (2013)。[29] Mohan Sridharan、Michael Gelfond、Shiqi Zhang 和 Jeremy Wyatt。2019。REBA：一种用于机器人学中知识表示与推理的基于细化的架构。人工智能研究杂志65 (2019), 87–180。[30] Eleonora Tagliabue、Daniele Meli、Diego Dall’Alba 和 Paolo Fiorini。2022。自主机器人手术中的审议：处理解剖结构不确定性的框架。在 *2022*年*IEEE*机器人与自动化国际会议（*ICRA 2022*）。[31] Yiming Wang、Francesco Giuliari、Riccardo Berra、Alberto Castellini、Alessio Del Bue、Alessandro Farinelli、Marco Cristani 和 Francesco Setti。2020。POMP：基于POMCP的在线运动规划，用于室内环境中的主动视觉搜索。在 英国机器视觉会议论文集（*BMVC*）。[32] Yue Wang、Abdullah Al Redwan Newaz、Juan David Hernández、Swarat Chaudhuri 和 Lydia E Kavraki。2021。针对具有安全可达性目标的部分可观测马尔可夫决策过程进行在线部分条件计划合成：方法与实验。*IEEE*自动化科学与工程汇刊18, 3 (2021), 932–945。[33] Maddalena Zuccotto、Marco Piccinelli、Alberto Castellini、Enrico Marchesini 和 Alessandro Farinelli。2022。在 POMCP 中学习状态变量关系：移动机器人的框架。机器人与人工智能前沿9 (2022)。