# 模仿学习方法综述及其在机器人领域的应用

李帅龙[1,2,3],张会文[1,2,3],周维佳[1,2]
1.中国科学院沈阳自动化研究所 机器人学国家重点实验室,沈阳 110016
2.中国科学院 机器人与智能制造创新研究院,沈阳 110016
3.中国科学院大学,北京 100049

**摘 要:**模仿学习一直是人工智能领域的研究热点。模仿学习是一种基于专家示教重建期望策略的方法。近年来,在理论研究中,此方法和强化学习等方法结合,已经取得了重要成果;在实际应用中,尤其是在机器人和其他智能体的复杂环境中,模仿学习取得了很好的效果。主要阐述了模仿学习在机器人学领域的研究与运用。介绍了和模仿学习相关的理论知识;研究了模仿学习的两类主要方法:行为克隆学习方法和逆强化学习方法;对模仿学习的成功应用进行总结;最后,给出当前面对的问题和挑战并且展望未来发展趋势。

**关键词:**人工智能;行为克隆;逆强化学习;模仿学习
**文献标志码:**A **中图分类号:**TP242.6 doi:10.3778/j.issn.1002-8331.1810-0007

李帅龙,张会文,周维佳.模仿学习方法综述及其在机器人领域的应用.计算机工程与应用,2019,55(4):17-30.
LI Shuailong, ZHANG Huiwen, ZHOU Weijia. Review of imitation learning methods and its application in robotics. Computer Engineering and Applications, 2019, 55(4):17-30.

## Review of Imitation Learning Methods and Its Application in Robotics

LI Shuailong[1,2,3], ZHANG Huiwen[1,2,3], ZHOU Weijia[1,2]
1.State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China
2.Institute for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110016, China
3.University of Chinese Academy of Sciences, Beijing 100049, China

**Abstract:** Imitation Learning(IL) has always been a research hotspot in the field of artificial intelligence. Imitation learning is a kind of method that reconstructs desired policies based on expert demonstrations. This method has been combined with methods such as reinforcement learning in theoretical research, and has achieved important results. In practical applications, especially in the complex environment of robots and other agents, imitation learning has achieved good results. This paper elaborates the research and application of imitation learning about robotics. Firstly, theoretical knowledge related to imitation learning is introduced. Secondly, two main methods of imitation learning are studied: Behavioral Cloning (BC) method and Inverse Reinforcement Learning(IRL) method. Thirdly, the successful applications of imitation learning are summarized. Finally, it is necessary to summarize the current problems and challenges and look forward to the future development trend.

**Key words:** artificial intelligence; behavioral cloning; inverse reinforcement learning; imitation learning

## 1 引言

在传统方式中,机器人和其他智能体系统以人工编程的方式控制其自主行为,需要专门的知识和技能。但是,机器人和其他智能体的应用环境已经从简单的环境转移到复杂的非结构化的环境,人工编程行为越来越具有挑战性,并且耗时、代价昂贵[1-2]。最重要的是,由于机器人行为的多样性,人类不可能编程机器人的所有行为。从示教中学习期望行为是一种很有效的解决方法,模仿学习(Imitation Learning,IL)就是基于这一思想而进行的工作。在IL中,专家示教提供了有效信息,从而

**作者简介:**李帅龙(1989—),男,博士研究生,研究领域为人工智能、模仿学习、机器学习;张会文(1991—),通讯作者,男, 博士研究生,研究领域为模仿学习、强化学习、机器学习,E-mail:zhanghuiwen@sia.cn;周维佳(1957—),男,博士,教授,研究领域为空间机器人学、空间自主有效载荷。

**收稿日期:**2018-10-28 **修回日期:**2018-12-07 **文章编号:**1002-8331(2019)04-0017-14

提高了学习效率,并且适用于复杂任务,而无需人工编程所需要的相关专业技能和知识。传感器技术的进步、计算能力的提高以及深度学习等相关科学研究的发展又一步提高了IL的性能,如实时性、鲁棒性和表示能力等。

目前,IL作为机器人学的一个重要分支,已经受到了研究者的广泛关注。其中,Autonomous Systems Labs利用IL方法分别对球杯运动、绳球运动和抓取目标等运动技能进行研究;Autonomous Motion Department利用IL方法对智能体的运动、感知和控制学习进行研究。近年来,OpenAI、DeepMind和Google Brain都分别开展了工作并取得了很好的效果。但是,国内研究者对IL的相关研究比较少。针对这一现状,本文对IL的相关研究和应用进行了全面的总结,包括IL的分类、方法以及研究成果和成功应用。

一般情况下,IL根据具体示教选择合适的算法,然后生成初始化策略,最后通过初始化策略和环境的交互不断地优化策略,以便得到期望策略。IL的一般流程如图1所示。IL可以分为两大类:一类称为行为克隆(Behavioral Cloning,BC),该方法直接从专家示教数据中学习期望策略;另一类称为逆强化学习(Inverse Reinforcement Learning,IRL),该方法利用恢复奖励函数方法间接地学习策略。

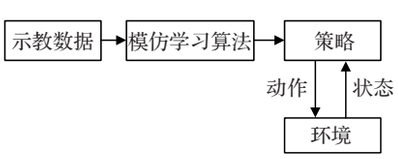


**图1　模仿学习的一般流程**

目前,IL方法已经应用于很多领域中,包括类人机器人[3-4]、游戏[5]、人机交互[6-7]、机器视觉[8-9]等,是人工智能的一个重要分支。本文对IL的研究历程、发展和应用进行详细阐述。本文的整体框架如图2所示。

## 2　预备知识

首先简要地介绍IL需要的相关知识,包括相关符号的表示、统计机器学习、强化学习和IL的形式化等。

### 2.1　符号

在引入机器学习的相关知识之前,首先介绍本文出现的一些符号。专家示教常常作为一组轨迹给出,轨迹的数据集为 $\mathcal{D}=\{\tau^0,\tau^1,\cdots,\tau^m\}$;使用 $q$ 代表专家策略概率分布;$p$ 代表学徒策略概率分布;$x$ 代表系统状态;$u$ 表示动作;$s$ 表示上下文;$T$ 代表有限时间步长,单个轨迹时间步的总数量为 $T+1$。其中,上下文 $s$ 代表不同的任务场景,可以是系统初始状态 $x_0$,或者相关对象的状态。表1总结了本文符号。

**表1　符号表**

| 符号 | 表示含义 |
|---|---|
| $x$ | 系统状态 |
| $s$ | 上下文 |
| $\boldsymbol{\Phi}$ | 特征向量 |
| $\tau$ | 轨迹 |
| $u$ | 动作 |
| $z$ | 观测 |
| $\pi^{\mathrm{E}}$ | 专家策略 |
| $\pi^{\mathrm{L}}$ | 学徒策略 |
| $\mathcal{D}$ | 示教数据集 |
| $q$ | 专家策略的概率分布 |
| $p$ | 学徒策略的概率分布 |
| $t$ | 时间步 |
| $T$ | 时间步长最大值 |
| $N$ | 示教数量 |
| $\tau^{\mathrm{demo}}$ | 专家示教轨迹 |

### 2.2　统计机器学习

#### 2.2.1　马尔科夫决策过程

在学习机器人的行为策略时,人们常常把机器人行为轨迹定义为一个马尔科夫决策过程。如果一个状态序列仅仅依赖 $x_{t+1},x_{t+2},\cdots$ 以往状态序列 $x_0,x_1,\cdots,x_t$ 中的 $x_t$,那么状态序列 $x_0,x_1,\cdots,x_t$ 是一个马尔科夫链。在马尔科夫链中,下一个状态 $x_{t+1}$ 只依赖于当前状态 $x_t$ 的性质称为马尔科夫性质。满足马尔科夫性质的过程称为马尔科夫决策过程。一个马尔科夫决策过程

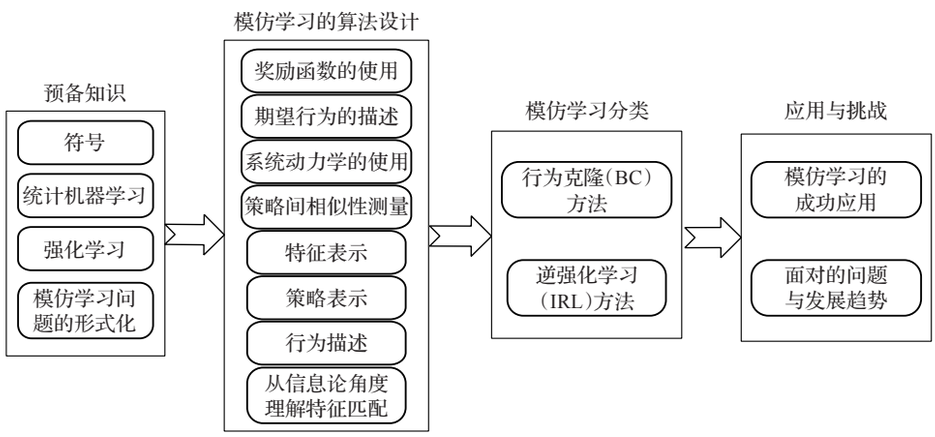


**图2　本文整体框架图**

可以表示为一个元组 $(X, U, P, \gamma, D, R)$。其中，$X$ 是状态的有限集；$U$ 是动作集；$P$ 是状态转移概率集；$\gamma$ 是折扣函数；$D$ 是由初始状态 $x_0$ 组成的初始状态分布；$R$ 是奖励函数。

#### 2.2.2　熵

机器人的状态转移是一个不确定量，用熵来表示机器人的不确定程度。给定状态变量 $x$ 和其概率分布 $p(x)$，它的熵为：

$$H(p)=-\int p(x)\ln p(x)\mathrm{d}x \tag{1}$$

熵是信息不确定程度的一种度量，且是一个凸函数。

#### 2.2.3　KL 散度

模仿学习需要比较学徒策略分布与专家策略分布的差异，以便于得到合适的学徒策略。在信息几何领域，KL（Kullback-Leibler）散度被用来量化两个概率分布的差异[10]，即：

$$D_{\mathrm{KL}}(p(x)\|q(x))=\int p(x)\ln\frac{p(x)}{q(x)}\mathrm{d}x \tag{2}$$

由上式可知，KL 散度可以测量两个概率分布的差异。另外，KL 散度是不对称的，即：

$$D_{\mathrm{KL}}(p(x)\|q(x))\neq D_{\mathrm{KL}}(q(x)\|p(x)) \tag{3}$$

KL 散度常常被用来度量 IL 方法[11]。

#### 2.2.4　信息与矩投影

模仿学习需要从数据集中学习期望策略，一个从数据集学习策略的通用方法是把数据集“投影”到策略模型的空间中。信息论使用两种投影：信息投影（I-projection）和矩投影（M-projection）[12]。使用 KL 散度[10]，信息投影为：

$$p^*=\arg\min_p D_{\mathrm{KL}}(p(x)\|q(x)) \tag{4}$$

矩投影为：

$$p^*=\arg\min_p D_{\mathrm{KL}}(q(x)\|p(x)) \tag{5}$$

因为 KL 散度是不对称的，所以两个投影的结果是不同的。

#### 2.2.5　最大熵原理

在给定专家示教下，模仿学习得到的解常常有很多，为了找到唯一解，需要一些额外约束。最大熵方法是一个有效的约束方法。

考虑一个学徒概率分布 $p(x)$，它与专家特征分布 $q(x)$ 匹配，满足：

$$E_p[\boldsymbol{\Phi}(x)]=E_q[\boldsymbol{\Phi}(x)] \tag{6}$$

其中，$E_q[\boldsymbol{\Phi}(x)]$ 是专家特征函数的期望；$E_p[\boldsymbol{\Phi}(x)]$ 是学徒特征函数的期望。由于这样的分布有很多，人们需要额外约束来获得唯一解[11]。例如，文献[13]就是选择最大熵的方法来求得唯一解：

$$H(p)=-\int p(x)\ln p(x)\mathrm{d}x \tag{7}$$

在约束优化过程中，最大熵分布可以视为下式：

$$p(x)\propto\exp(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\Phi}(x)) \tag{8}$$

其中，$\boldsymbol{\omega}$ 是用于特征匹配约束的拉格朗日乘子向量。从上式可以看出，在给定特征向量下，最大熵分布就可以匹配具体的特征期望。文献[14]把最大熵概念推广为最大因果熵。

### 2.3　强化学习

模仿学习和强化学习有密切的联系，尤其是在 IRL 方法中，因此有必要简要介绍强化学习知识。强化学习是一种从环境状态映射到动作的学习，目的是使得智能体在与环境交互过程中获得最大的奖励。强化学习的一般过程如图 3 所示。

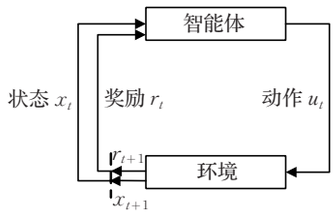


**图 3　强化学习的一般过程**

强化学习的目标是学习一个策略 $\pi$，这一策略的目标是把系统状态映射到控制输入中，以便于最大化期望奖励 $J(\pi)$。奖励函数 $r(t)$ 代表时间 $t$ 时给定状态、动作或者轨迹的性质。对于有限时间步 $T$，期望回报由每一时间步的奖励累积给出：

$$J(\pi)=E[\sum_{t=0}^{T}r(t)|T] \tag{9}$$

在无限步长时，可以引入折扣因子 $\gamma$：

$$J(\pi)=E[\sum_{t=0}^{T}\gamma^t r(t)|T] \tag{10}$$

其中，折扣因子 $\gamma$ 决定短期回报和长期回报之间的权衡。期望策略 $\pi^*$ 为：

$$\pi^*=\arg\max_{\pi}J(\pi) \tag{11}$$

在策略 $\pi$ 的状态值 $x$ 下，期望回报为：

$$V^{\pi}(x)=E[\sum_{t=0}^{\infty}r_t|x_0=x,\pi] \tag{12}$$

其中，$V^{\pi}(x_t)$ 称为值函数[15]。相似地，在策略为 $\pi$，状态值为 $x$，采取动作 $u$ 下的期望回报为：

$$Q^{\pi}(x,u)=E[\sum_{t=0}^{\infty}r_t|x_0=x,u_0=u,\pi] \tag{13}$$

### 2.4　模仿学习问题的形式化

IL 的目标是学习一个策略，这一策略可以重现专家示教执行的期望任务。假设专家示教的行为可以观察到一个轨迹为 $\tau=[\boldsymbol{\Phi}_0,\boldsymbol{\Phi}_1,\cdots,\boldsymbol{\Phi}_T]$，这是特征 $\boldsymbol{\Phi}$ 的一个序列。特征 $\boldsymbol{\Phi}$ 可以是机器人系统的状态或者其他测量，可以根据给定问题选择。

通常，记录示教的条件是不同的，例如在不同的位置抓取目标。这里把任务条件称为任务的上下文 $s$，并

和特征轨迹存储在一起。上下文 $s$ 可包含和任务相关的任何信息,例如机器人的初始状态或者任务目标的位置。但是,在执行期间上下文 $s$ 是固定的,唯一动态是状态特征 $\boldsymbol{\Phi}_t$。如有必要,奖励信号 $r$ 也可以和上下文 $s$、轨迹 $\tau$ 存储在一起。

在IL中,搜集示教数据集 $\mathcal{D}=(\tau_i,s_i,r_i)_{i=1}^N$,这一数据集由轨迹 $\tau$、上下文 $s$ 和可选择的奖励函数 $r$ 组成。其中,数据收集过程可以是在线或者离线的。利用收集的数据集 $\mathcal{D}$,一个通用的基于优化的策略方法为:

$$\pi^*=\arg\min_{\pi} D(q(\boldsymbol{\Phi}),p(\boldsymbol{\Phi})) \tag{14}$$

其中,$q(\boldsymbol{\Phi})$ 是专家策略的特征分布;$p(\boldsymbol{\Phi})$ 是学徒的特征分布;$D(q(\boldsymbol{\Phi}),p(\boldsymbol{\Phi}))$ 是相似性测量。当数据集包含多任务示教且上下文 $s$ 包括每个任务的相关信息时,这个问题可以认为是多任务学习[16-17]。

另外,人们常常使用一个环境,它可以交互地执行和评估策略,例如模拟器或者实体机器人系统。为了匹配示教,这样的环境可以收集新的数据并迭代地改善策略。

## 3　模仿学习的算法设计

当开发IL方法时,为了把问题规范化,明确设计选择是很有必要的。本文论述一些必要的设计选择。

### 3.1　奖励函数的使用

是否使用奖励函数是IL涉及的基本问题,并且涉及到与强化学习的界定问题。相对于强化学习,IL主要特征是可以使用专家示教,并且利用专家示教优化策略。IL是以专家信号为最优目标进行优化学习,以此提高学徒的学习水平。因此,区别于强化学习,在IL中,专家示教可以引导学徒进行有效学习,以避免大量的并且昂贵的全局探索。这样,IL就可以通过示教加速对特定问题的学习过程,并且样本数量呈现指数级改善[18]。

另一方面,IL是否使用奖励函数由实际的任务决定。如果人为定义奖励函数很困难,那么使用奖励函数是不切实际的[6,19]。

IL的通用方法是先利用示教信息初始化策略,然后通过不断试错的强化学习进行优化策略[20]。

### 3.2　期望行为的描述:行为克隆和逆强化学习

为了有效地学习数据,需要确定一个最有效的策略表示。一种策略的最有效表示常常是状态特征到轨迹/动作的直接映射,这一方法称为行为克隆(BC)。然而,对于集中于长期规划的问题,策略的最有效表示是把策略作为优化或者规划问题的方法[21-22]。逆强化学习(IRL)可以解决这一优化问题,其通过学习一个代价函数可以很好地学习专家示教。BC是一种在给定的状态-动作对和上下文的示教数据集 $\mathcal{D}=(\tau_i,r_i,s_i)_{i=1}^N$ 的情况下,利用监督学习得到的由状态和上下文到动作的方法。而IRL是一种从示教中恢复回报函数的方法。在IRL中,专家示教是最优的行为,通过 $\pi^*=\arg\max_{\pi} J(\pi)$ 来从专家示教中恢复奖励函数。$J(\pi)$ 是给定策略 $\pi$ 下累积回报的期望。

从BC和IRL的定义可以看出,两类方法的选择要依赖具体问题。若从状态特征到动作的映射是可行的,就使用BC方法;若可以设计一个回报函数并且能很好地表示期望行为,则可以使用IRL方法。因此,分析期望行为如何执行很重要。

### 3.3　系统动力学的使用:基于模型(model-based)和无模型(model-free)

为了使一些问题易于处理,需要使用系统动力学。例如,对于欠驱动机器人的运动规划问题,由于可达到状态有限,建立系统动力学模型规划可行轨迹是必要的。对于给定的控制系统,确定是否使用系统动力学很重要。在IL中,把学习一个系统动力学模型并利用它学习策略的方法称为基于模型方法,而不能明确表示并学习系统动力学模型的方法称为无模型方法。基于模型和无模型方法的优缺点如表2所示。

**表2　无模型和基于模型方法在模仿学习中的优缺点**

| 特点 | 基于模型 | 无模型 |
|---|---|---|
| 优点 | 学习过程提高数据利用率;基于系统动力学 | 无需学习系统动力学 |
| 缺点 | 学习系统动力学困难 | 不能保证在给定系统中使用 |

一般情况下,BC主要集中于无模型方法,IRL主要集中于基于模型方法。出现这种情况是有原因的。对于机器人系统,行为克隆主要集中于行为轨迹规划,系统动力学并不是关键因素,因此无模型是一个很好的选择。而IRL是学习一个策略,这一策略的状态-动作空间需要在给定系统中迭代求值,在系统动力学已知的情况下,学徒的表现更容易预测,因此基于模型方法很适合。但是系统动力学的学习具有挑战性。

### 3.4　特征表示

选择合适的并且有能力表达期望行为的特征很重要。在限制学习复杂度的同时,特征还需包含解决问题的足够信息。特征可以是和期望任务相关的多种度量。

### 3.5　策略表示

好的策略表示可以更好地得到期望行为。策略表示与任务的抽象水平有关,像任务水平、轨迹水平和状态-动作水平,人们需要确定学习什么水平的任务。为建立期望行为的模型,有效的策略表示是必要的,但是策略表示复杂度的增加往往会导致训练数据和训练时间的增加。

人们把策略抽象表示分为三个水平:任务水平抽象、轨迹水平抽象和动作-状态水平抽象。在任务水平抽象中,学徒可以学习一个策略,这一策略可以生成一

个选择 $o\in O$, $O$ 是可供选择的集合。供选择的集合 $O$ 通常定义为在一段时间内采取行动的策略。在任务水平的抽象中,每一个选择可以表示为动作集或轨迹,从给定状态 $x_t$ 和上下文 $s$ 到选择序列的映射策略可以表示为:

$$\pi:\boldsymbol{x}_t,s\rightarrow[o_1,o_2,\cdots,o_T] \tag{15}$$

其中, $T$ 为任务步长。任务水平抽象能够将这样复杂的任务建模为一系列简单的动作。

在轨迹水平的规划中,策略是从上下文 $s$ 到系统状态序列轨迹 $\tau$ 的映射:

$$\pi:s\rightarrow\tau \tag{16}$$

在状态-动作水平,策略是系统状态 $x_t$ 和上下文 $s$ 到动作 $u_t$ 的映射:

$$\pi:\boldsymbol{x}_t,s\rightarrow u_t \tag{17}$$

### 3.6 策略间相似性测量

在没有奖励函数的清晰概念时,需要建立专家策略和学徒策略之间的相似性重现专家行为。虽然相似性概念优先定义在学徒和系统共同定义的轨迹上,但是这种相似性也可以在个体决策的水平上建立。

### 3.7 行为描述

在IL中,量化行为并测量专家行为和学徒行为之间的不同很重要。给定一个由状态-动作对组成的数据集 $D=(\boldsymbol{x},u)$ ,人们可以建立状态-动作对的联合分布 $p(\boldsymbol{x},u)$ 或者给定状态对应动作的条件概率分布 $p(u|\boldsymbol{x})$ 。早期的IL方法通过监督学习建立状态-动作对来学习策略。但是,因为状态-动作对分布仅仅描述短期行为,所以匹配状态-动作对分布会导致长期行为的误匹配,这种误匹配称为级联错误。

为了匹配专家和学徒的长期行为,考虑轨迹特征是有必要的。因为轨迹或者轨迹的观察常常是随机的并且带有噪音,所以轨迹特征的期望被用来描述专家和学徒的行为。学徒策略的轨迹特征期望为:

$$E_{p(\tau)}[\boldsymbol{\Phi}(\tau)]=\int p(\tau)\boldsymbol{\Phi}(\tau)\mathrm{d}\tau \tag{18}$$

其中, $p(\tau)$ 是学徒策略的轨迹分布; $\boldsymbol{\Phi}(\tau)$ 是轨迹 $\tau$ 的特征向量。当轨迹 $\mathcal{D}=\{\tau_i\}_{i=1}^N$ 的数据集可用,轨迹特征的期望可以近似表示为:

$$E_{p(\tau)}[\boldsymbol{\Phi}(\tau)]\simeq\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\Phi}(\tau_i^{\mathrm{demo}}) \tag{19}$$

匹配特征期望在BC和IRL中都有出现。轨迹特征的分布 $p(\boldsymbol{\Phi}(\tau))$ 常常被用来匹配专家和学徒的行为特征。人们不仅可以匹配一阶矩,还可以匹配高阶矩。轨迹分布 $p(\tau)$ 可以认为是特征分布的特例。

### 3.8 从信息论角度理解特征匹配

IL可以描述为寻求策略问题,这一策略是示教和学徒差异的最小解。为此,很多IL方法把示教行为映射到参数化策略空间。把示教投影到参数化策略需要考虑示教分布和参数化策略分布的关系。信息论提供了评估这一关系的方法。假设一个轨迹分布为 $p(\tau|\boldsymbol{\omega})$ ,这一分布是由带有参数向量 $\boldsymbol{\omega}$ 的策略引导的。监督学习方法常常是基于给定数据的最大似然来求解。以M-投影为例:

$$\boldsymbol{\omega}^*=\arg\min_{\boldsymbol{\omega}}D_{\mathrm{KL}}(q(\tau)\|p(\tau|\boldsymbol{\omega})) \tag{20}$$

其中, $q(\tau)$ 是专家策略引导轨迹的经验分布, $\tau$ 表示一个轨迹, $p(\tau|\boldsymbol{\omega})$ 是学徒轨迹的概率分布。由此产生的目标函数可以写为:

$$\begin{aligned}\Gamma&=D_{\mathrm{KL}}(q(\tau)\|p(\tau|\boldsymbol{\omega}))=\int q(\tau)\ln\frac{q(\tau)}{p(\tau|\boldsymbol{\omega})}\mathrm{d}\tau=\\&E_q[\ln q(\tau)]-E_q[\ln p(\tau|\boldsymbol{\omega})]\end{aligned} \tag{21}$$

其中, $E_q$ 是关于 $q(\tau)$ 的期望,用于评估示教轨迹。因上式的第一项独立于 $\boldsymbol{\omega}$,所以通过最大化第二项 $E_q[\ln p(\tau|\boldsymbol{\omega})]$ 可以最小化 $D_{\mathrm{KL}}(q(\tau)\|p(\tau|\boldsymbol{\omega}))$ 。基于简单监督学习的IL可以视为计算M-投影的特殊情况,因为这些算法在本质上是执行似然最大化。值得注意的是,极大似然解与最大熵原理的解之间存在着密切的关系。

从以上所述可以看出,这些设计选择是相互关联的、灵活的。比如策略的选择与模型选择、特征表示等有直接关系。

## 4 行为克隆

BC方法是学习一个从状态/上下文到轨迹/动作的映射,而无需求解奖励函数。当这种映射是表示期望策略的最有效方式时,BC方法是再现示教行为的有效方法。

机器人系统的控制器可以看成分层结构。上层结构的控制器基于给定的上下文和观测来规划期望轨迹;下层结构的控制器控制动作完成期望轨迹。对于机器人系统,BC的主要目标是学习这些控制器。

对于上层结构,BC的目标是学习一个可以生成期望轨迹的策略。其中,上下文 $s$ 可以是机器人操作系统的初始状态 $x_0$ 或者与给定任务相关的对象状态。对于下层结构,专家示教的数据集为 $\mathcal{D}=(u_i,s_i)_{i=1}^N$ 。IL的目标是在给定状态 $x_t$ 和上下文 $s$ 的情况下学习一个可以生成动作的策略。从以上分析可以看出,IL方法是监督学习问题,策略可以通过一个简单的回归问题得到。IL的目标是学习一个特定应用中的策略 $\pi_\theta$ 。IL的通用算法如算法1所示。

**算法1** 行为克隆方法

收集专家示教轨迹 $D$

选择策略表示 $\pi_\theta$

选择目标函数 $\Gamma$

基于示教 $\mathcal{D}$ ,调整策略参数 $\theta$ 优化 $\Gamma$

返回 优化的策略参数 $\theta$

其中,目标函数 $\Gamma$ 是表示专家示教和学徒策略的相似性;$\theta$ 为待调整的策略参数。

从算法1可以看出,在给定示教的情况下,选择合适的策略表示和目标函数很重要。两者的选择最终决定了算法及其性能。

目标函数一般表示示教策略和学徒策略的相似性,可以用损失函数来表示。IL使用的常见损失函数为二次损失函数、L1-损失函数、L2-损失函数[23]、KL散度[24]、对数损失函数[23]等。

### 4.1　克隆方法的选择

选择BC方法,合适的策略表示是重要的。比如,简单的策略表示易于训练但是信息量少;复杂的策略表示需要大量训练数据。下面分别针对基于模型方法和无模型方法进行讨论。其中,无模型分别从任务水平、轨迹水平和状态-动作对水平三方面进行讨论,基于模型则主要讨论使用前向动力学的模仿学习和基于迭代学习控制的模仿学习。

### 4.2　基于无模型的行为克隆方法

无模型的行为克隆方法无需学习系统动力学并且无需恢复奖励函数进行学习重现专家行为的策略。与基于模型方法相比,无模型方法无需迭代地学习,执行时相对简单。但是学习得到的轨迹不能保证在给定系统中使用,因此,无模型很难应用到可达状态集受限的欠驱动系统中。

#### 4.2.1　状态-动作对水平的无模型行为克隆方法

早期的IL研究把无模型BC方法理解为监督学习[7,19,25]。文献[19]训练一个神经网络用于自动驾驶系统,这一神经网络建立一个从摄像机图像到转向角映射的模型。但是这一工作在实践中并不成功。主要原因有两种:一是由于示教数据集有限,学徒遇到的状态分布与给定的示教数据集状态分布不同,而监督学习是基于训练数据集样本是独立同分布的假设,因此监督学习很难泛化到未知场景;二是不可避免的级联错误得不到纠正。IL方法违背独立同分布原理,而在一定规模的IL问题中,收集所有可能情况下的示教是不可行的;在级联错误中,每一步微小错误累积为大错误,导致学徒与专家示教有很大偏差,使得到的性能不佳。为解决这两个问题,示教-学徒交互式学习策略应运而生。交互式学习策略可以在策略更新和当前状态分布之间交替性学习。

结构化预测方法是提出由输入 $x$ 到输出 $y$ 的映射函数[26-27]。文献[28]提出基于搜索的结构化预测算法(SEARN)。在此算法中,结构化预测问题规约为简单的分类问题,结构化预测作为结构化输出 $y$ 的分量 $y_t$ 的搜索过程,并且第 $t$ 步的决策依赖于 $t-1$ 步决策。因此,分类器的训练过程依赖于分类器自身。结构化预测算法与支持向量机和条件随机场相比具有更快的学习能力[28]。

文献[29]提出一个基于置信度的方法,在给定状态的置信度学习策略时,此方法基于置信度确定是否需要额外的专家示教。通过能返回置信度的分类器,学徒决定怎样从动作集中选择动作。当置信度低于阈值,就需要额外的专家示教。通过额外的示教,此算法试图在学徒策略的诱导下学习策略。当专家观察到学习者的不正确动作时,专家纠正动作,并将校正的动作添加到训练数据集。但是,在置信度方法中,阈值直接影响方法的效果,这是该方法的难点,动作标注的修改也是该方法的难点。

文献[30]提出一种数据聚合(Data Aggregation Approach,DAGGER)算法。此算法通过学徒策略诱导下的状态分布来收集专家示教。主要思想是每一次迭代都使用专家策略 $\pi^{\mathrm{E}}$ 和学徒策略 $\pi_i^{\mathrm{L}}$ 来获得数据 $\mathcal{D}_i^{\mathrm{L}}$,然后把数据 $\mathcal{D}_i$ 并入原数据集 $\mathcal{D}$ 得到新的数据集 $\mathcal{D}$,根据数据集 $\mathcal{D}$ 训练策略 $\pi_{i+1}^{\mathrm{L}}$,最后返回最好的验证策略。其中数据集 $\mathcal{D}^i$ 的状态由专家策略 $\pi^{\mathrm{E}}$ 和学徒策略 $\pi_i^{\mathrm{L}}$ 共同提供,动作由专家示教提供。这一算法的优势是每一次迭代都会增加新的数据 $\mathcal{D}^i$ 并重新训练分类器,可以使学习器及时地从错误中恢复过来。DAGGER算法如算法2所示。DAGGER算法对于简单问题和复杂问题都适用,仅需很少的迭代训练,并且数据越多,效果越好。此算法的困难点是需要对动作进行重新标记。

**算法2**　DAGGER算法

初始化 $\mathcal{D}\leftarrow\varnothing,\pi_1\leftarrow\pi^*$

For　$i=1:N$

利用 $\pi_i$ 为步长为 $T$ 的轨迹采样

得到数据集 $\mathcal{D}^i=\{(x,\pi^*(s))\}$,其中状态 $x$ 由学徒策略 $\pi_i^{\mathrm{L}}$ 得到,动作 $u$ 由专家策略 $\pi^{\mathrm{E}}$ 给出

聚合数据 $\mathcal{D}=\mathcal{D}\bigcup\mathcal{D}_i$

在数据集 $\mathcal{D}$ 上训练策略 $\pi_{i+1}^{\mathrm{L}}$

返回　在验证集运行最好的策略 $\pi_i^{\mathrm{L}}$

文献[5]基于DAGGER算法提出了DaD(Data as Demonstrator)算法。此算法把多步预测问题理解为IL问题,多步预测的级联错误通过数据聚合方法进行改善。

#### 4.2.2　轨迹水平的无模型行为克隆方法

在机器人操作中,轨迹规划是最重要的问题之一。如果假设系统全驱动并且低级控制器可以实现期望的状态,那么无需评估系统动力学就可以学习给定任务的轨迹。因为商业机器人操作系统经常使用这样的低级控制器,所以对于机器人操作系统轨迹规划的IL研究,无模型的BC方法是一个重要领域。

给定一个轨迹数据集 $\mathcal{D}$,可以学习一个直接从上下文映射到轨迹 $\tau$ 的策略,如式(16)。为此,可以使用机器学习中的多种回归方法,例如高斯混合回归方法[31]

和高斯过程回归[6]。但是在机器人操作系统中并不仅限于回归方法。为保证在机器人系统中的适用性,需要在规划的轨迹中加入一些约束,比如平滑地收敛于目标状态。为学习这些策略,回归方法需要满足期望的约束。例如,文献[32]结合交叉熵优化(Cross Entropy Optimization,CEO)方法和高斯混合回归(Gaussian Mixture Regression,GMR)方法,即交叉熵回归(Cross Entropy Regression,CER)方法,提高了算法的泛化能力。

在计算机图形学领域,基于关键帧(Keyframe)方法和基于关键点(Via-Point)方法用于表示实现给定任务的重要状态。任务轨迹的状态在关键帧方法中表示为一系列关键帧,在关键点方法中表示为关键点。文献[33]运用关键帧方法从人类专家示教中学习跳舞。通过三维运动,跟踪系统进行捕捉人类专家的运动,提取关键帧。根据仿人系统动力学修改关键帧,仿人系统可以执行示教舞蹈。文献[34]通过学习音乐和关键帧的依赖关系学习到了跟随音乐跳舞的策略。此方法的困难点是修改关键帧/关键点。

隐马尔科夫模型(Hidden Markov Models,HMMs)是常常用于建立离散状态之间概率转换的模型。HMMs由一个有限隐状态集 $X$ 、有限观察标签集 $Y$ 、状态转移矩阵 $A$ 、输出概率矩阵 $B$ 和一个初始分布 $d_i$ 组成。给定观察序列和状态集,通过Baum-Welch算法(期望最大化(EM)算法的变型)求得 $A$ 和 $B$ ,进而可以求得给定初始状态下的运动序列。HMMs利用学习的概率模型识别当前的系统状态,并用于示教轨迹的聚类和分割中[35-37]。HMMs的缺点是表示的离散性。状态数量多会导致计算成本过高,状态数量少不能有效表示轨迹。为了克服这一缺点,此模型和其他算法结合,比如用状态高斯模型表示像速度、空间位置和力等连续值[38]。最近提出的HMMs用来建立一个复杂的状态持续分布[39]。文献[40]利用LQR控制器处理从隐半马尔科夫模型(Hidden Semi-Markov Models,HSMMs)中检索的轨迹优化问题。文献[41]提出基于HMMs的关节力矩规划方法来控制接触力。

动态运动基元(Dynamic Movement Primitives,DMPs)是在文献[42]中提出的。动态运动基元是吸引子动力学微分方程驱动的,其表示的示教运动是非线性力项与吸引子力项的结合。非线性力可以表示复杂运动,吸引子力代表目标状态。非线性力随着时间弱化,最后吸引子力占主导地位,因此动态运动基元可以平滑地收敛到目标状态。DMPs保证了轨迹的平稳性和连续性,并且可以表示非线性运动而不失稳定性。DMPs表示是一种确定性方法表示运动,而专家示教常常是随机的,因此文献[43]提出了概率运动基元(Probabilistic Movement Primitives,ProMPs)。ProMPs表示轨迹的分布,但是不能保证规划轨迹的稳定性。

#### 4.2.3 任务水平的无模型行为克隆方法

(1)分割和聚类

当一个任务需要一个复杂的运动,这一运动称为运动基元序列。这种高水平运动规划称为任务水平规划[44]。

虽然无模型的轨迹学习常常隐含地假设每一个示教包含单个运动基元,但是在实践中每个示教轨迹由不同类型的运动基元组成。因此,为了学习每一个运动基元,分割示教轨迹是有必要的。另外,为了学习运动基元的多种类型,把分割的运动进行聚类也是很有必要的。但是人工分割和聚类是费时的,因此在模仿学习领域已经研究了关于分割和聚类的方法。

文献[45]开发了基于HMMs的在线分割方法。此方法通过计算附近数据窗口之间的距离,运用监督学习来分割运动数据。文献[46]提出了一种使用分解HMMs的分割和聚类方法。此方法计算HMMs之间的距离,观测的运动分割聚类为一个树结构。文献[47]提出的 $\beta$ 过程自回归隐马尔科夫模型(Beta Process Autoregressive HMMs)是一个贝叶斯非参数化方法,目的是为了找到时间序列数据的动态特征。文献[48]利用 $\beta$ 过程自回归马尔科夫方法来学习机器人的运动基元序列。先前的研究中,IL中轨迹分割的进展和机器学习领域的方法论进展密切相关。

(2)学习运动基元序列

为了学习运动基元序列,为技能结构建模和学习示教行为的运动基元之间的转变是有必要的。

学习运动基元的一种方法是学习一种树状的技能结构。文献[49]提出一个在线学习算法从示教中构建技能树,并且应用于机器人的路径规划中。

另外一种序列运动的方法是学习不同运动基元之间的转变模型。文献[50]学习运动基元库并且使用支持向量机来计算当前的运动基元到下一个运动基元的多分类问题的解。

为了学习运动基元之间的概率转换模型,基于HMMs的方法被应用。在经典的自回归隐马尔科夫模型(Autoregressive Hidden Markov Models,AR-HMMs)中,当前的状态仅仅依赖于先前状态;与其相反,文献[51]提出的自回归隐马尔科夫模型(State-based Transitions Autoregressive Hidden Markov Models,STARHMMs),其隐变量的概率分布依赖于观察状态,隐变量用来表示任务的当前阶段。文献[52]的框架使用自回归隐马尔科夫模型,把任务表示为一个确定性运动基元(DMP)序列,其中变量表示当前的激活DMP。该模型使用条件运动基元规划,这一规划可以基于观察把一个DMP转化为另一个DMP。

实际应用中,示教数据常常是不足的。为了处理这一问题,文献[48]提出了面向任务水平规划的增量IL方法。其框架利用非结构化示教和人类专家的纠正行

为。文献[48]应用 $\beta$ 过程自回归隐马尔科夫模型分割示教任务,把离散基元之间的转化作为有限状态自动机(Finite State Automaton,FSA)并建立模型。当给出一个新的状态时,学徒利用训练的FSA把任务规划为一个运动基元序列。如果专家认为有必要细化规划运动,可以停止运动的执行并纠正动作。因此,在此方法中学徒通过与专家示教的交互来改善性质。

利用示教运动的注释来学习任务水平规划也是一个有趣的方法。文献[37]开发了一种方法来学习语言和标注的互动模式,利用示教运动数据集和标注语句。在这个框架中,一个运动语言模型通过潜在变量学习运动符号和单词之间的关系,一个自然语言模型运用 $n$ 元语法模型学习语句结构。表3总结了本文中无模型行为克隆主要方法的优点和局限性。

### 4.3 基于模型的行为克隆方法

在模仿学习中,智能体试图模仿专家示教。如前所述,对于欠驱动系统,由于可达状态受限,当规划可行的轨迹时往往需要考虑系统动力学。模仿学习的一个问题是"对应问题"[1,53]。对应问题是指学徒的实施例、速度等和示教之间的匹配问题。学徒和示教之间在规模和结构上的任何差异都需要在训练时得到补偿。"对应问题"也需要系统动力学调整规划轨迹。基于模型的IL方法就是通过学习系统动力学来试图重现示教行为。

#### 4.3.1 基于前向动力学的行为克隆方法

解决"对应问题"的直接方式是学习一个系统的前向动力学模型。轨迹规划基于此前向动力学模型来规划轨迹。前向动力学模型可以视为回归问题。基于模型方法的早期研究中,经常使用局部权重回归和高斯混合回归;近期研究使用高斯过程,可以模拟任何黑盒函数,并且可以模拟不确定性。但是从高斯过程原始方程中可以看出,计算复杂度会随着数据点的增加而立方增加,即使当前最优的近似算法中,也会成平方增加。由于前向动力学模型是学习下一状态与当前状态-动作对(假设维数为 $n$ )的关系,若利用高斯过程学习前向动力学模型,随着时间步长 $t$ 的增加,高斯过程的协方差矩阵大小为 $tn\times tn$ ,计算复杂度变为 $O((tn)^3)$ ,导致计算成本显著增大,实时性得不到保证。因此高斯过程不适合高维数据,也不适合时间步长很大的情况。深度学习的方法[9,17,34,54]被用来处理高维数据。高斯混合模型(Gaussian Mixture Models,GMMs)[55]和高斯过程(Gaussian Processes,GPs)[56]用来预测前向动力学。文献[57]仅仅利用观察进行学习,利用逆动力学模型学习动作并建立动力学模型,然后从示教中学习策略。文献[9]利用深度学习网络从图像中学习逆动力学模型,然后利用示教做高水平的引导并和低水平的逆动力学模型共同完成任务。但是,深度学习网络最大缺点是需要大量示教数据集。

#### 4.3.2 基于迭代学习控制的行为克隆方法

为了开发一个控制器来完成期望轨迹,也可以无需前向动力学模型而使用一个迭代学习控制方法。文献[58]使用二次线性调节器(Linear Quadratic Regulator,LQR)迭代地学习一个控制器来重建期望轨迹。该方法简单易行,但不能推广到不同的期望轨迹。文献[59]提出迭代的二次线性调节器(Iteration Linear Quadratic Regulator,iLQR),通过迭代学习过程学习线性反馈控制器以跟随轨迹,常常用于动力学并不准确的系统中。

相对于基于模型的BC方法,无模型的BC方法因为不需学习系统动力学,所以不需迭代学习,容易实现。但是,在学习轨迹时,无模型BC并不能保证在给定系统中所学轨迹的适用性。为此,难以将无模型方法应用于可达到状态集受限的欠驱动系统中。与无模型BC方法相反,基于模型的BC方法应用系统动力学的信息学习策略。即使在欠驱动的状态下,通过学习前向模型,基于模型的行为克隆方法也可以找到近似于专家行为的合适的轨迹。但是,学习前向模型不是一件容易的工作,并且基于模型的BC方法常常需要迭代学习,时间

**表3 无模型行为克隆方法的优点和局限性**

| 方法 | 优点 | 局限性和缺点 |
|---|---|---|
| 监督学习方法 | 方法简单,易实现 | 监督学习的样本是独立同分布假设,难于泛化;易于发生级联错误 |
| 结构化预测方法 | 与支持向量机、条件随机场等相比,有更快的学习能力 | 分类器的训练过程依赖于分类器自身,而非示教 |
| 基于置信度方法 | 是示教-学徒交互式学习方法,训练集中不断添加新的专家示教 | 合适的阈值选择;修改动作标注 |
| 数据聚合 | 适合于简单问题和复杂问题;需少量的迭代训练;数据越多,效果越好;级联错误得到改善 | 需要对动作进行重新标注 |
| 高斯过程 | 可以模拟任何函数;可以模拟不确定性 | 随着时间步长的增加,计算复杂度显著增大 |
| 关键帧/关键点方法 | 直接使用计算机视觉进行示教 | 关键帧/关键点标注的修改 |
| 隐马尔可夫模型 | 对过程的预测效果良好 | 表示是离散的,状态数量少不能有效表示轨迹 |
| 动态运动基元 | 可以表示复杂运动;保证了轨迹的平稳性和连续性,并且不失稳定性 | 是一种确定性方法,而专家示教常常是随机的 |
| 概率运动基元 | 用随机分布表示轨迹的分布 | 不能保证轨迹的稳定性 |

成本和计算成本较大。表4总结了本文中基于模型行为克隆方法的优点和局限性。

**表4 基于模型行为克隆方法的优点和局限性**

| 方法 | 优点 | 局限性和缺点 |
|---|---|---|
| 前向动力学模型 | 有效地解决了对应问题 | 学习前向动力学有困难 |
| 迭代学习控制方法 | 简单易行;应用于动力学不准确的系统中;无需学习前向动力学模型 | 需要迭代学习;时间成本和计算成本大 |

## 5 逆强化学习

IRL试图从示教策略中恢复奖励函数。当奖励函数是描述期望行为的最有效方式时,恢复奖励函数是有效的。在IRL中,一个通用假设是示教者利用马尔科夫决策过程进行决策。另外,很多IRL方法假设存在特征向量 $\boldsymbol{\Phi}:X\rightarrow[0,1]^K$ 。IRL方法常常把奖励函数视为关于这些特征的函数。

IRL的目标是从专家轨迹中恢复未知奖励函数 $R(\tau)$ 。然而,一个最优策略可以有多个奖励函数,这一问题称为不适定问题。在IRL中为了获得唯一解,很多研究提出优化附加目标函数,像最优策略和其他策略之间的余量[60]、最大化熵[61]方法。很多算法通常需要一个迭代学习过程。算法3总结了一类通过交替地求解一个强化学习类型问题和更新成本函数估计IRL方法。为了获得与专家策略相同的性质,状态-动作对访问频率 $\mu$ 需要在示教轨迹和学徒策略引导的轨迹匹配[62]。通过在期望特征匹配约束下优化目标函数,更新奖励函数参数 $\omega$ 。设计目标函数来评估奖励函数,使得示教策略比当前策略更理想。使用一种基于当前评估的奖励函数的优化控制解法(例如RL),更新策略参数 $\theta$ 。为了这一目的,IRL方法在内循环中常常有一个强化学习过程。通过重复这一过程,可以获取策略和奖励函数参数。

**算法3** 特征匹配逆强化学习的抽象版本

输入 示教轨迹 $\mathcal{D}=(\boldsymbol{\tau}_i)_{i=1}^N$
初始化奖励函数和策略参数 $\boldsymbol{\omega},\theta$
重复 在当前策略 $\boldsymbol{\pi}_\theta$ 下评估状态动作对的访问频率 $\mu$
评估目标函数 $\Gamma$ 及其导数 $\nabla_\omega\Gamma$
更新奖励函数参数 $\boldsymbol{\omega}$
用强化学习方法更新策略参数 $\theta$
返回 策略参数 $\theta$ 和奖励函数参数 $\boldsymbol{\omega}$

每一个IRL算法都有一个不同的方法来执行这些步骤。为了评估状态-动作对访问频率 $\mu$ ,基于模型方法需要系统动力学知识,而无模型方法则利用基于采样的方法。为了获得基于恢复奖励函数的最优策略,可以使用多种RL方法。例如,马尔科夫决策过程(Markov Decision Process,MDP)用于离散的状态-动作对空间的策略优化[63],文献[64]利用了引导策略搜索,文献[62,65]利用了信赖域策略优化。

像BC方法一样,IRL方法也可以分为两类:基于模型方法和无模型方法。基于模型的IRL方法假设系统动力学是已知的。系统动力学的先验知识常常用来评估和更新所学的奖励函数或策略。当系统动力学已知时,基于模型的IRL方法可以很简单地执行。然而,由于非线性动力学很难评估,其应用于非线性动力学具有挑战性。另一方面,无模型的IRL方法不需要系统动力学的先验知识,而是使用基于样本的方法评估和更新学习的奖励函数或策略,并且可以应用到非线性动力学系统。但是多次采样来评估轨迹分布,导致这种方法耗时并且计算量很大。

### 5.1 基于模型的逆强化学习方法

本节回顾基于IRL方法,其利用系统动力学的先验知识。

文献[66]为解决IRL问题提出了匹配特征期望,提出从示教中学习策略,以最大化最优策略和其他策略的不同。通过迭代地更新学习策略,此算法可以找到与示教策略相近的最优策略。

匹配特征期望也在其他IRL方法中出现[67]。因为多个策略可以达到相同的期望特征计数,所以匹配特征计数有多解。因此,满足最优策略附加条件是必要的。

为获得唯一解,文献[63]提出最大间隔规划(Maximum Margin Planning,MMP)。MMP是找到最优策略和其他策略最大差异的成本函数。基于MMP,文献[60,68]提出LEARCH(Learning to Search)方法。与MMP相比,此方法中成本函数为非线性,利用指数函数梯度下降法来优化最大间隔规划目标。MMP方法是生成确定性优化策略,但是对于大空间维数的机器人在规划中常常需要随机性策略和近似策略。最大熵IRL可以很好地解决这一问题。

在最近的IRL研究中,最大熵原理常常被用来获得唯一奖励函数。虽然MMP在有明显优于备选方案的单个奖励函数上效果很好,但是在优化其他案例中的行为分布时效果不理想。最大熵理论旨在选择一个分布,这一分布匹配示教特征期望的分布并且熵最大。根据这一理论,文献[67]提出学习一个最大化熵方法,但是仅仅适用于确定性环境。对于随机性环境,IRL最大熵方法的主要理论缺陷之一是存在关于随机环境中转移概率的有偏项。IRL的最大因果熵方法[69]可以解决这一问题。

最大因果熵的关键思想是动作选择是因果的,即在时间步 $t$ 的动作选择独立于轨迹的未来状态。与最大熵方法相反,IRL的最大因果熵方法去除了由于环境的随机性动力学引起的有偏项。

文献[61]把IRL的最大因果熵方法扩展到失败示教中,进而从失败示教中学习策略。文献[61]考虑学徒与失败示教的特征差异性,并且融入到IRL的最大因果熵方法[69],利用梯度下降法得出特征权重。这种方法的关

键点首先要找到成功示教的权重,然后集中于找到失败示教的权重。此方法因为需要学习失败示教的权重,所以增加了方法的复杂度。

近期的IRL研究中,最大熵理论是IL的主要研究方向,但是也有很多其他基于模型的IRL方法。文献[70]提出的线性可解的马尔科夫决策过程(MDP)方法不同于通用的IRL方法,其评估的是值函数而不是奖励函数或者成本函数。文献[70]从一个无约束凸优化问题找到一个极大似然值函数。此方法的优点是不需要重复地求解MDP。缺点是在连续状态空间中近似求解值函数是很有挑战性的;运用IRL通用方法近似求解得到的奖励函数通用性不强,很难泛化到其他的应用场景中。

基于贝叶斯框架的IRL方法是由文献[4]提出的。这一框架中,专家的行为是更新先前奖励函数的证据,并且计算后验概率均值需要恢复奖励函数和从示教中学习优化策略。在文献[4]中,应用MCMC(Markov Chain Monte Carlo)算法从分布中生成样本,样本均值是真实分布均值的一个估计。文献[71]不计算后验均值,而是提出最大后验推断(Maximum A Posteriori,MAP)。

原始的IRL方法主要集中于学习与特征向量线性相关的奖励函数,下面讨论基于模型的非线性奖励。

文献[63]使用的是增强方法,这些算法可以通过组合监督学习算法来创造更好的非线性成本函数。但是多种算法同时学习,计算量增加。文献[68]运用此方法从示教中学习运动策略。文献[72-73]提出IRL成本函数的深度神经网络方法。此方法基于最大熵方法,并且使用反馈变量从示教中学习复杂的成本函数。文献[74]提出使用基于最大熵原理的高斯过程方法,IRL的高斯过程方法中的奖励函数关于特征是非线性的。文献[64]基于非线性成本函数方法[72],使用自适应采样方法的策略优化步骤。最近,生成对抗网络[75]已经引入到IRL中[64-65]。在生成对抗网络中,一个生成模型 $G$ 训练一个用于模仿真实数据分布的生成数据样本,而判别器 $D$ 用于判别数据是否是真实数据。这一工作表明优化过程扮演生成器的角色,学习的成本函数作为判别器的角色。生成对抗模仿学习框架图如图4所示。文献[54]把生成对抗模仿学习方法扩展到基于模型方法中,提出使用一个前向模型来训练一个完全可微分的随机策略。实验结果表明,在连续的控制任务中,基于模型的生成对抗模仿学习的性能优于无模型的生成对抗模仿学习。但是,由于一些情况下生成对抗网络收敛性差而难以训练,模型过于自由、难于控制也是此算法的难点。文献[76]仅从观察示教中学习策略,先建立逆动力学模型,然后从示教中学习策略。表5总结了本文中基于模型逆强化学习方法的优点和局限性。

## 5.2　无模型的逆强化学习方法

在机器人系统和其他一些应用中,人们很难得到精确的动力学模型,而基于模型的方法往往需要系统动力学。因此就需要一种方法来避开求解系统动力学问题。无模型的IRL方法是基于样本的方法来评估轨迹分布,虽然在学习过程中需要很多轨迹样本,但是避免了系统动力学的学习过程。

文献[77]提出相对熵方法,通过最小化一个基准策略的先验轨迹分布 $q_0(\tau)$ 和学徒策略的轨迹分布 $p(\tau)$ 的

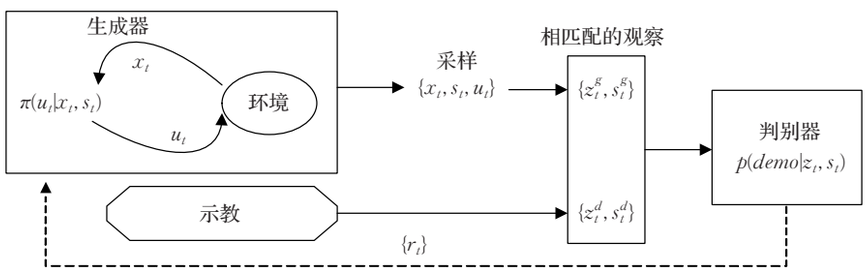


**图4　生成对抗模仿学习框架图**

**表5　基于模型逆强化学习方法的优点和局限性**

| 方法 | 优点 | 局限性和缺点 |
|---|---|---|
| 最大间隔规划方法 | 有效地解决了不适定问题 | 生成的是确定性优化策略,不能使用于随机性策略;不适用于最优策略和备选方案差别不明显的情况 |
| 最大熵原理 | 适用于最优策略和备选方案不明显的情况 | 仅适用于确定性环境,对于随机性策略,存在随机环境中转移概率的有偏项 |
| 最大因果熵方法 | 去除了环境中随机动力学引起的有偏项 | 与特征向量是线性相关;难以应用于非线性场景 |
| 马尔科夫过程方法 | 从无约束凸优化问题找到极大似然值函数;不需重复求解MDP | 在连续状态空间中求解值函数有困难;通用性不强 |
| 增强学习方法 | 组合监督学习算法学习非线性成本函数 | 多种算法同时学习,计算量大 |
| 生成对抗网络 | 无需预先建模;不需遵循任何因式分解的模型,任何生成器网络和鉴别器都起作用 | 难以训练,收敛性差;模型过于自由、不可控 |

相对熵来求得轨迹分布,进而得到奖励函数中的参数。此方法利用重要性采样估计特征期望,因此不需要知道系统动力学。

文献[65]提出生成对抗模仿学习,把生成对抗网络[75]和IRL结合。该方法能够根据未知的奖励函数来约束智能体行为到近似最优,而无需明确地尝试恢复该奖励函数。文献[65]训练重现专家行为策略的生成器和区分学徒策略轨迹和专家示教策略轨迹的判别器,并且使用信赖域策略优化方法[78]来优化目标函数。为了避免策略更新的不稳定性,信赖域策略优化方法中使用KL散度测量策略之间的差异性。文献[79]利用生成对抗网络从无结构化、无标签的示教中建立多模态模仿学习框架。文献[80-81]把生成对抗模仿学习扩展到分层策略的框架中。文献[80]利用生成对抗模仿学习从运动捕捉数据集中学习低水平控制器的具体技能,利用强化学习使智能体和环境交互学习高水平控制器。文献[80]只需要无示教动作 $u$ 的部分动作特征,并且无需大量的专业知识。但是,实验结果显示由于运动捕捉示教数据集有噪音和智能体自身因素,实验结果不理想。options框架可以把策略分为子策略,文献[81]扩展了此框架,把奖励函数也扩展为对应的子奖励函数,提出了奖励-策略options框架,利用生产对抗逆强化学习方法学习options。此方法可以应用在简单环境和复杂环境中。

## 6 模仿学习的应用

本文简要阐述IL在机器人领域的相关应用,分别从BC和IRL两大类别,是否基于模型进行介绍。

### 6.1 无模型行为克隆方法的相关应用

文献[42]利用确定性马尔科夫决策过程学习网球挥拍运动。示教数据是由穿戴运动捕捉服的人提供。为准确地重现轨迹,实验中运用了逆动力学控制器。实验结果表明学习得到的运动可以应用到不同的目标位置。文献[82]运用运动觉示教进行模仿球拍动作使球在球拍上反复弹跳。

文献[6]在机器人手术中学习打结任务(Knot-Tying)。在给定上下文的情况下,其利用高斯过程学习示教轨迹下的条件分布,并且可以把示教轨迹推广到新的上下文中。学习得到的轨迹分布也可以应用于规划和控制手术器械和物体之间的接触力。

文献[3]把随机性马尔科夫过程(ProMPS)和IL方法相结合,学习人机交互中的交接任务。

### 6.2 基于模型行为克隆方法的相关应用

文献[30]利用DAGGER算法学习自动飞行任务,有效地改善了级联错误中示教与学徒之间的偏差。文献[24]利用高斯过程学习前向模型,把运动觉示教应用于欠驱动机器人中,相对于无模型方法具有更强的鲁棒性。文献[9]利用自我监督学习成功地完成了机器人对变形物体(绳子)的操作。

### 6.3 基于模型逆强化学习方法的相关应用

文献[77]应用相对熵方法学习球杯运动(ball-in-a-cup),成本函数用随机梯度下降法获得,并取得好的效果。文献[66]学习模仿不同的驾驶风格。

### 6.4 无模型逆强化学习方法的相关应用

文献[60]利用LEARCH算法学习特征函数的代价,并且通过经典运动规划方法找到了优化路径。文献[64]利用神经网络学习非线性成本函数,并且成功应用到机器人操作中。文献[72]把MMP方法应用到自动驾驶中。

## 7 未来的挑战与发展趋势

尽管IL方法表现出很强的能力,但是仍然存在一些挑战与问题。

### 7.1 面对的挑战与问题

处理示教数据问题是最先遇到的挑战。示教数据的处理直接关系到算法选择、处理难度和性能好坏。目前遇到的有关问题包括如何处理示教中不良运动,如何从原始传感器输入的示教中学习,如何利用相关任务示教学习当前任务。

怎么学的问题是另一个重要的挑战,其中包括相似性测量,从多种传感器数据中学习,如何把示教和先验知识结合,如何选择轨迹表示,如何学习示教不能做的任务以及算法选择问题。

性能评估也是IL的一个挑战。因为IL的应用很广泛,所以相应的性能评估也是一个开放性问题。主要问题包括如何建立IL标准和评估IL的度量方法。

### 7.2 模仿学习的发展趋势

IL作为人工智能的研究方向之一,易于应用在工程实际中,已经吸引了学术界和工业界人士的兴趣并得到持续的研究与发展。在这篇综述中,讨论了IL的相关知识、方法和一些实际应用。为了进一步提高IL的通用性和工程实际应用,未来的IL会有如下的几个发展方向:(1)趋向于有效且自动地抽取特征。(2)任务学习方法研究趋向于少样本方法研究和通用性方法研究。(3)目前主要集中于单个智能体IL方法研究,但是随着应用环境复杂度的提高,多智能体系统的协调控制有着广泛的应用场景,受到国内外学者的广泛关注。虽然关于多智能体系统的研究层出不穷,但是仍然没有形成普遍使用的理论和方法,并且此方向还有很多问题有待研究,例如群集问题、一致性问题和网络优化问题。因此学习多智能体之间的协作和交互会是未来发展趋势。(4)更加趋向于通过增量式学习方式来训练IL模型。(5)把IL应用到无法看到的场景是一个重要发展方向。(6)多模态IL将会提高IL的效果,是一个重要的发

展方向。可以预见,随着IL理论和方法研究的深入,人类将会更好地利用IL解决实际问题。

## 参考文献:


[1] Hussein A,Gaber M M,Elyan E,et al.Imitation learning:a survey of learning methods[J].ACM Computing Surveys,2017,50(2):1-35.

[2] Osa T,Pajarinen J,Neumann G,et al.An algorithmic perspective on imitation learning[J].Foundations and Trends® in Robotics,2018,7(1/2):1-179.

[3] Maeda G J,Neumann G,Ewerton M,et al.Probabilistic movement primitives for coordination of multiple human-robot collaborative tasks[J].Autonomous Robots,2017,41(3):593-612.

[4] Ramachandran D,Amir E.Bayesian inverse reinforcement learning[C]//International Joint Conference on Artificial Intelligence,2007:2586-2591.

[5] Venkatraman A,Hebert M,Bagnell J A.Improving multi-step prediction of learned time series models[C]//29th AAAI Conference on Artificial Intelligence,2015:3024-3030.

[6] Osa T,Sugita N,Mitsuishi M.Online trajectory planning and force control for automation of surgical tasks[J].IEEE Transactions on Automation Science and Engineering,2018,15(2):675-691.

[7] Chambers R A,Michie D.Man-machine co-operation on a learning task[M]//Computer graphics.Boston,MA:Springer,1969:179-186.

[8] Sermanet P,Xu K,Levine S.Unsupervised perceptual rewards for imitation learning[J].arXiv:1612.06699,2016.

[9] Nair A,Chen D,Agrawal P,et al.Combining self-supervised learning and imitation for vision-based rope manipulation[C]//2017 IEEE International Conference on Robotics and Automation,2017:2146-2153.

[10] Kullback S,Leibler R A.On information and sufficiency[J].The Annals of Mathematical Statistics,1951,22(1):79-86.

[11] Amari S.Information geometry and its applications[M].Berlin:Springer,2016.

[12] Bishop C M.Graphical models[J].Pattern Recognition and Machine Learning,2006,4:359-422.

[13] Jaynes E T.Information theory and statistical mechanics[J].Physical Review,1957,106(4):620.

[14] Ziebart B D,Bagnell J A,Dey A K.The principle of maximum causal entropy for estimating interacting processes[J].IEEE Transactions on Information Theory,2013,59(4):1966-1980.

[15] Sutton R S,Barto A G.Reinforcement learning:an introduction[M].Cambridge:MIT Press,1998.

[16] Duan Y,Andrychowicz M,Stadie B,et al.One-shot imitation learning[C]//Advances in Neural Information Processing Systems,2017:1087-1098.

[17] Finn C,Abbeel P,Levine S.Model-agnostic meta-learning for fast adaptation of deep networks[J].arXiv:1703.03400,2017.

[18] Sun W,Venkatraman A,Gordon G J,et al.Deeply aggrevated:differentiable imitation learning for sequential prediction[J].arXiv:1703.01030,2017.

[19] Pomerleau D A.ALVINN:an autonomous land vehicle in a neural network[C]//Advances in Neural Information Processing Systems,1989:305-313.

[20] Silver D,Huang A,Maddison C J,et al.Mastering the game of Go with deep neural networks and tree search[J].Nature,2016,529:484-489.

[21] Ratliff N D,Silver D,Bagnell J A.Learning to search:functional gradient techniques for imitation learning[J].Autonomous Robots,2009,27(1):25-53.

[22] Bagnell J A.An invitation to imitation[R].Carnegie Mellon University,Robotics Institute,2015.

[23] Sugiyama M.Introduction to statistical machine learning[M].[S.l.]:Morgan Kaufmann,2015.

[24] Englert P,Paraschos A,Deisenroth M P,et al.Probabilistic model-based imitation learning[J].Adaptive Behavior,2013,21(5):388-403.

[25] Widrow B,Smith F W.Pattern-recognizing control systems[C]//Proceedings of Computer and Information Science Symposium,1964.

[26] Tsochantaridis I,Joachims T,Hofmann T,et al.Large margin methods for structured and interdependent output variables[J].Journal of Machine Learning Research,2005,6:1453-1484.

[27] Bakir G,Hofmann T,Schölkopf B,et al.Predicting structured data[M].Cambridge:MIT Press,2007.

[28] Daumé H,Langford J,Marcu D.Search-based structured prediction[J].Machine Learning,2009,75(3):297-325.

[29] Chernova S,Veloso M.Interactive policy learning through confidence-based autonomy[J].Journal of Artificial Intelligence Research,2009,34:1-25.

[30] Ross S,Gordon G,Bagnell D.A reduction of imitation learning and structured prediction to no-regret online learning[C]//Proceedings of the 14th International Conference on Artificial Intelligence and Statistics,2011:627-635.

[31] Calinon S,Guenter F,Billard A.On learning,representing,and generalizing a task in a humanoid robot[J].IEEE Transactions on Systems,Man,and Cybernetics:Part B,2007,37(2):286-298.

[32] 张会文,张伟,周维佳.基于交叉熵优化的高斯混合模型运动编码[J].机器人,2018,40(4):569-576.

[33] Nakaoka S, Nakazawa A, Kanehiro F, et al.Learning from observation paradigm: leg task models for enabling a biped humanoid robot to imitate human dances[J]. The International Journal of Robotics Research, 2007, 26(8): 829-844.

[34] Okamoto T, Shiratori T, Kudoh S, et al.Toward a dancing robot with listening capability: keypose-based integration of lower-, middle-, and upper-body motions for varying music tempos[J].IEEE Transactions on Robotics, 2014, 30(3): 771-778.

[35] Lee D, Ott C, Nakamura Y.Mimetic communication model with compliant physical contact in human—humanoid interaction[J].The International Journal of Robotics Research, 2010, 29(13): 1684-1704.

[36] Takano W, Nakamura Y.Real-time unsupervised segmentation of human whole-body motion and its application to humanoid robot acquisition of motion symbols[J]. Robotics and Autonomous Systems, 2016, 75: 260-272.

[37] Takano W, Nakamura Y.Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions[J].The International Journal of Robotics Research, 2015, 34(10): 1314-1328.

[38] Racca M, Pajarinen J, Montebelli A, et al.Learning in-contact control strategies from demonstration[C]//2016 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2016: 688-695.

[39] Calinon S, Pistillo A, Caldwell D G.Encoding the time and space constraints of a task in explicit-duration hidden Markov model[C]//2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2011: 3413-3418.

[40] Rozo L, Silverio J, Calinon S, et al.Learning controllers for reactive and proactive behaviors in human-robot collaboration[J].Frontiers in Robotics and AI, 2016, 3: 30.

[41] Takano W, Nakamura Y.Planning of goal-oriented motion from stochastic motion primitives and optimal controlling of joint torques in whole-body[J].Robotics and Autonomous Systems, 2017, 91: 226-233.

[42] Ijspeert A J, Nakanishi J, Schaal S.Learning attractor landscapes for learning motor primitives[C]//Advances in Neural Information Processing Systems, 2003: 1547-1554.

[43] Paraschos A, Daniel C, Peters J R, et al.Probabilistic movement primitives[C]//Advances in Neural Information Processing Systems, 2013: 2616-2624.

[44] Lagriffoul F, Dimitrov D, Bidot J, et al.Efficiently combining task and motion planning using geometric constraints[J].The International Journal of Robotics Research, 2014, 33(14): 1726-1747.

[45] Kohlmorgen J, Lemm S.A dynamic hmm for on-line segmentation of sequential data[C]//Advances in Neural Information Processing Systems, 2002: 793-800.

[46] Kulić D, Takano W, Nakamura Y.Incremental learning, clustering and hierarchy formation of whole body motion patterns using adaptive hidden Markov chains[J]. The International Journal of Robotics Research, 2008, 27(7): 761-784.

[47] Fox E, Jordan M I, Sudderth E B, et al.Sharing features among dynamical systems with beta processes[C]// Advances in Neural Information Processing Systems, 2009: 549-557.

[48] Niekum S, Osentoski S, Konidaris G, et al.Learning grounded finite-state representations from unstructured demonstrations[J].The International Journal of Robotics Research, 2015, 34(2): 131-157.

[49] Konidaris G, Kuindersma S, Grupen R, et al.Robot learning from demonstration by constructing skill trees[J].The International Journal of Robotics Research, 2012, 31(3): 360-375.

[50] Manschitz S, Kober J, Gienger M, et al.Learning movement primitive attractor goals and sequential skills from kinesthetic demonstrations[J].Robotics and Autonomous Systems, 2015, 74: 97-107.

[51] Kroemer O, Van Hoof H, Neumann G, et al.Learning to predict phases of manipulation tasks as hidden states[C]// 2014 IEEE International Conference on Robotics and Automation, 2014: 4009-4014.

[52] Kroemer O, Daniel C, Neumann G, et al.Towards learning hierarchical skills for multi-phase manipulation tasks[C]// 2015 IEEE International Conference on Robotics and Automation, 2015: 1503-1510.

[53] Billard A, Calinon S, Dillmann R, et al.Robot programming by demonstration[M]//Springer handbook of robotics. Berlin, Heidelberg: Springer, 2008: 1371-1394.

[54] Baram N, Anschel O, Caspi I, et al.End-to-end differentiable adversarial imitation learning[C]//International Conference on Machine Learning, 2017: 390-399.

[55] Creating brain-like intelligence: from basic principles to complex intelligent systems[M].Berlin: Springer, 2009.

[56] Deisenroth M P, Englert P, Peters J, et al.Multi-task policy search for robotics[C]//2014 IEEE International Conference on Robotics and Automation, 2014.

[57] Torabi F, Warnell G, Stone P.Behavioral cloning from observation[J].arXiv: 1805.01954, 2018.

[58] Abbeel P, Coates A, Ng A Y.Autonomous helicopter aerobatics through apprenticeship learning[J].The International Journal of Robotics Research, 2010, 29(13): 1608-1639.

[59] Tassa Y,Erez T,Todorov E.Synthesis and stabilization of complex behaviors through online trajectory optimization[C]//2012 IEEE/RSJ International Conference on Intelligent Robots and Systems,2012:4906-4913.

[60] Silver D,Bagnell J A,Stentz A.Learning from demonstration for autonomous navigation in complex unstructured terrain[J].The International Journal of Robotics Research,2010,29(12):1565-1592.

[61] Shiarlis K,Messias J,Whiteson S.Inverse reinforcement learning from failure[C]//Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems,2016:1060-1068.

[62] Ho J,Gupta J,Ermon S.Model-free imitation learning with policy optimization[C]//International Conference on Machine Learning,2016:2760-2769.

[63] Ratliff N D,Bagnell J A,Zinkevich M A.Maximum margin planning[C]//Proceedings of the 23rd International Conference on Machine Learning,2006:729-736.

[64] Finn C,Levine S,Abbeel P.Guided cost learning:deep inverse optimal control via policy optimization[C]//International Conference on Machine Learning,2016:49-58.

[65] Ho J,Ermon S.Generative adversarial imitation learning[C]//Advances in Neural Information Processing Systems,2016:4565-4573.

[66] Abbeel P,Ng A Y.Apprenticeship learning via inverse reinforcement learning[C]//Proceedings of the 21st International Conference on Machine Learning,2004:1.

[67] Ziebart B D,Maas A L,Bagnell J A,et al.Maximum entropy inverse reinforcement learning[C]//Proceedings of the 23rd National Conference on Artificial Intelligence,2008:1433-1438.

[68] Zucker M,Ratliff N,Stolle M,et al.Optimization and learning for rough terrain legged locomotion[J].The International Journal of Robotics Research,2011,30(2):175-191.

[69] Ziebart B D.Modeling purposeful adaptive behavior with the principle of maximum causal entropy[D].Pittsburgh:Carnegie Mellon University,2010.

[70] Dvijotham K,Todorov E.Inverse optimal control with linearly-solvable MDPs[C]//Proceedings of the 27th International Conference on Machine Learning,2010:335-342.

[71] Choi J,Kim K E.Map inference for Bayesian inverse reinforcement learning[C]//Advances in Neural Information Processing Systems,2011:1989-1997.

[72] Grubb A,Bagnell J A.Boosted backpropagation learning for training deep modular networks[C]//Proceedings of the 27th International Conference on Machine Learning,2010:407-414.

[73] Bradley D M.Learning in modular systems[R].Carnegie Mellon University,Robotics Institute,2010.

[74] Levine S,Popovic Z,Koltun V.Nonlinear inverse reinforcement learning with Gaussian processes[C]//Advances in Neural Information Processing Systems,2011:19-27.

[75] Goodfellow I,Pouget-Abadie J,Mirza M,et al.Generative adversarial nets[C]//Advances in Neural Information Processing Systems,2014:2672-2680.

[76] Edwards A D,Sahni H,Schroeker Y,et al.Imitating latent policies from observation[J].arXiv:1805.07914,2018.

[77] Boularias A,Kober J,Peters J.Relative entropy inverse reinforcement learning[C]//Proceedings of the 14th International Conference on Artificial Intelligence and Statistics,2011:182-189.

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

[79] Hausman K,Chebotar Y,Schaal S,et al.Multi-modal imitation learning from unstructured demonstrations using generative adversarial nets[C]//Advances in Neural Information Processing Systems,2017:1235-1245.

[80] Merel J,Tassa Y,Srinivasan S,et al.Learning human behaviors from motion capture by adversarial imitation[J].arXiv:1707.02201,2017.

[81] Henderson P,Chang W D,Bacon P L,et al.Optiongan:learning joint reward-policy options using generative adversarial inverse reinforcement learning[J].arXiv:1709.06683,2017.

[82] Kober J,Peters J.Learning motor primitives for robotics[C]//2009 IEEE International Conference on Robotics and Automation,2009:2112-2118.