计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20170578
56(2):254-261, 2019

基于逆强化学习的示教学习方法综述

张凯峰　　俞　扬
(计算机软件新技术国家重点实验室(南京大学)　南京　210023)
(zhangkf@lamda.nju.edu.cn)

Methodologies for Imitation Learning via Inverse Reinforcement Learning: A Review

Zhang Kaifeng and Yu Yang
(*State Key Laboratory for Novel Software Technology* (*Nanjing University*), *Nanjing* 210023)

Abstract　Motivated by applying reinforcement learning methods into autonomous robotic systems and complex decision making problems, reinforcement learning is becoming more and more popular in the community of machine learning. Traditional reinforcement learning is one kind of learning paradigm in machine learning field which is learning from the interactions between the agent and the environment. However, for the vast majority of cases, the environments for sequential decision making problems cannot provide an explicit reward signal immediately or the reward signal can be much delayed. This becomes the bottleneck for applying reinforcement learning methods into more complex tasks. So inverse reinforcement learning is proposed to recover the reward function from expert demonstrations in the Markov decision process (MDP) by assuming that the expert demonstrations is optimal. So far, the imitation learning algorithms which combines direct reinforcement learning approaches and inverse reinforcement learning approaches have already made a great progress. This paper briefly introduces the basic concepts of reinforcement learning, inverse reinforcement learning and imitation learning. And this paper also gives an introduction to the existing problems concerning with inverse reinforcement learning and some other methods in imitation learning. In addition, we also introduce some existing bottlenecks once applying the above methods into real world applications.

Key words　reinforcement learning; imitation learning; inverse reinforcement learning; Markov decision process; multi-step decision problem

摘　要　随着强化学习在自动机器人控制、复杂决策问题上的广泛应用,强化学习逐渐成为机器学习领域中的一大研究热点.传统强化学习算法是一种通过不断与所处环境进行自主交互并从中得到策略的学习方式.然而,大多数多步决策问题难以给出传统强化学习所需要的反馈信号.这逐渐成为强化学习在更多复杂问题中实现应用的瓶颈.逆强化学习是基于专家决策轨迹最优的假设,在马尔可夫决策过程中逆向求解反馈函数的一类算法.目前,通过将逆强化学习和传统正向强化学习相结合设计的一类示教学习算法已经在机器人控制等领域取得了一系列成果.对强化学习、逆强化学习以及示教学习方法做一定介绍,此外还介绍了逆强化学习在应用过程中所需要解决的问题以及基于逆强化学习的示教学习方法.

关键词　强化学习;示教学习;逆强化学习;马尔可夫决策过程;多步决策问题

中图法分类号　TP391

收稿日期:2017-08-11;**修回日期**:2018-06-05

基金项目:江苏省自然科学基金项目(BK20160066)
This work was supported by the Natural Science Foundation of Jiangsu Province (BK20160066).

通信作者:俞扬(yuy@nju.edu.cn)

强化学习(reinforcement learning, RL)[1]是机器学习的重要分支之一.在强化学习中,智能体(agent)通过不断与其所处环境(environment)自主交互从而进行学习并完成任务.在交互过程中,智能体将基于最大化累积反馈奖赏的目标对自身策略不断进行优化更新.该过程可以被认为是(正向)强化学习过程.与传统监督学习不同的是,强化学习天生具有一定的"自学"能力,可以自主地对环境进行探索学习.因此,强化学习能够被有效地应用到许多标记数据代价高昂的自主学习问题当中去,这包括:推荐系统、自动驾驶、智能机器人、Atari 游戏等.在强化学习中,如图 1 所示,智能体通过观测所处环境的状态,在动作空间选取合适动作予以执行.环境将依据相应状态转换概率转换至新的状态,并给予智能体一定反馈奖赏.这个过程可以始终执行下去,也可以在智能体观测到终止状态后停止.

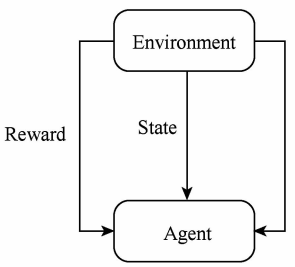

Fig. 1 Reinforcement learning procedure

图 1 强化学习过程

然而对于绝大多数决策问题而言,环境将难以给出准确的即时反馈信号或者环境给出的反馈信号将具有很高的延迟性.例如在自动驾驶问题中,对于行驶过程中的车辆,环境很难在车辆每执行一个动作后即时地给出反馈信号;而在围棋这一类游戏之中,在每一步的落子后环境也很难立即评价该步的好坏,而往往需要经过多步之后才能来判断之前一步的好坏,这也就是环境所给予的反馈信号延迟性较高的情况.在上述情况下,更为直接的方式是利用大量人类专家的决策数据进行学习从而得到智能体的策略.这样的学习方式被称为示教学习或模仿学习(imitation learning)[2].

示教学习的目标是模仿专家的决策轨迹进行决策,其中每条专家决策轨迹$\{\zeta_1,\zeta_2,\cdots,\zeta_m\}$包括了一系列的状态-动作对$\zeta_i=\langle s_{i1},a_{i1},s_{i2},a_{i2},\cdots,s_{in},a_{in}\rangle$.近年来,示教学习先后通过学习人类飞行员的飞行操作数据、道路导航数据以及自动系统控制数据等,在 Stanford 自动直升机[3-8]、导航[9-12]以及 HVAC 控制[13]等项目中取得了一系列成果.

根据模拟专家行为的不同实现过程,示教学习可以被划分为以下 3 种实现方式:

1) 行为克隆(behavioral cloning)[14-15].通过传统监督学习方法建立状态-动作之间的分类模型(针对离散动作空间)或回归模型(针对连续动作空间),从而实现决策,也即动作的预测.然而,由于该类方法在大规模状态空间下所得到的策略存在严重的复合误差(compounding errors)[16]并且难以有效学习到专家决策行为的动机.因此,该类方法需要设计人工标记数据的方法进行矫正,例如 DAgger 等[17],且仅适用于状态空间较小的情况.

2) 基于逆强化学习的示教学习方法.逆强化学习的目标是通过在马尔可夫决策过程上建立合适的优化模型,逆向求解得到决策问题的反馈函数.通过结合传统的正向强化学习方法设计的一系列示教学习方法,例如学徒学习(apprenticeship learning)[18]、代价指导学习(guided cost learning)[19]等,能够更好地解决大规模状态空间所带来的问题,因而在众多机器人项目中得到了广泛的应用.值得说明的是,部分研究工作也认为逆强化学习是示教学习方法的一种[20],这是由于该类方法在工作过程中通过不断的正向策略搜索进而优化算法所需要的反馈信号,因此整个系统(逆强化学习)可以被认为是一类示教学习方法.

3) 基于博弈的示教学习方法.经典的示教学习过程可以看作是智能体和所处环境进行博弈的过程.其中系统依据其混合策略P_t在动作空间选取动作,环境依据相应混合策略Q_t选取状态,同时系统将观测到自身在执行决策之后所得到的损失值.相关的经典工作包括通过已有自适应博弈方法[21]来优化学徒学习的 MWAL 算法[22],以及生成式对抗性示教学习方法[20,23-24],通过生成器(generator)生成策略,由判别器(discriminator)判断其是否是来自专家决策数据抑或是生成器生成的策略数据,通过训练 2 个学习器,寻找最优策略.

1 基本概念

在强化学习中,马尔可夫决策过程[1]可以形式化为一个五元组$\langle S,A,T,R,\gamma\rangle$表示.其中,$S$表示强化学习智能体所处环境的状态空间;$A$表示智能体可选取动作的动作空间;$T$表示状态转换概率模型;$R$表示环境在某个状态-动作对下所给予的反馈信号;$\gamma$表示反馈奖赏折扣系数.通常,强化学习

所面对的任务的状态转换模型以及反馈量需要通过智能体不断地探索(exploration)从而获取相关信息.

智能体的目标是通过和环境的不断交互最大化自身策略的未来累计反馈奖赏值.其交互过程为:智能体在某个状态 s_0 出发,根据策略在动作空间选取动作 a_1 执行,此时环境将依据其状态转换模型转换到下一个状态,同时将给予智能体一个确定的反馈奖赏.该过程将不断进行直到终止状态.其中智能体的策略 π 是指状态空间到动作空间的映射.

1.1 值函数

与动态规划算法类似的是,我们可以为每个状态定义一个值函数(value function),这将为强化学习的实现带来很大方便.值函数根据其自变量的不同可以分为:状态值函数 $V(s)$ 和状态-动作对值函数 $Q(s,a)$.其表述形式分别为

$$V^{\pi}(s)=E_{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t}r_{t}\,|\,s_{0}=s\right], \tag{1}$$

$$Q^{\pi}(s,a)=E_{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t}r_{t}\,|\,s_{0}=s,a_{0}=a\right]. \tag{2}$$

可以看出:状态值函数或者状态-动作对值函数分别是某个状态、状态-动作对下的累计未来反馈奖赏.因此只需要通过最大化值函数就可以最大化累计反馈奖赏,这使得强化学习策略求解更加方便.

基于最优策略,我们不难得到以下 2 个定理:

定理 1. Bellman 等式.假设马尔可夫决策过程为 $M=\langle S,A,T,R,\gamma\rangle$,智能体策略为 $\pi:S\to A$,对于任意状态 s、动作 a,其价值函数可以表示为

$$V^{\pi}(s)=R(s)+\gamma\sum_{s'\in S}T_{s\pi(s)}(s')V^{\pi}(s'), \tag{3}$$

$$Q^{\pi}(s,a)=R(s)+\gamma\sum_{s'\in S}T_{sa}(s')V^{\pi}(s'). \tag{4}$$

定理 2. Bellman 最优定理.假设马尔可夫决策过程为 $M=\langle S,A,T,R,\gamma\rangle$,智能体策略为 $\pi:S\to A$,则策略 π 是最优策略当且仅当对任意状态 s:

$$\pi(s)\in\arg\max_{a\in A}Q^{\pi}(s,a). \tag{5}$$

1.2 策略搜索

经典的正向强化学习研究是智能体基于最大化累计未来反馈奖赏求解策略的过程.而求解策略可以通过求解值函数实现.

根据 1.1 节所述,求解值函数可以通过式(6)和式(7)展开进行:

$$V^{*}(s)=\max_{a\in A}E[r_{t+1}+\gamma V^{*}(s_{t+1})\,|\,s_{t}=s,a_{t}=a]=\max_{a\in A}\sum_{s'\in S}T(s,a,s')[R(s,a)+\gamma V^{*}(s')], \tag{6}$$

$$Q^{*}(s,a)=E[r_{t+1}+\gamma\max_{a'\in A}Q^{*}(s_{t+1},a')\,|\,s_{t}=s,a_{t}=a]=\sum_{s'\in S}T(s,a,s')[R(s,a)+\gamma\max_{a\in A}Q^{*}(s',a)]. \tag{7}$$

通过式(6)(7)求解值函数从而获得最优策略的方法可以理解为策略迭代过程,也即通过不断迭代以下 2 个交互过程:策略评估(policy evaluation)和策略改进(policy improvement),从而获取最优策略.其中,策略评估是指通过当前的策略评估值函数,而策略改进是指通过当前值函数优化得到新的策略.这个过程就是经典的正向强化学习过程.

2 逆强化学习

逆强化学习是通过大量专家决策数据在马尔可夫决策过程中逆向求解环境反馈信号函数的一类方法.其基本原则是寻找一个或多个反馈信号函数能够很好地描述专家决策行为.这也就是说,逆强化学习算法将基于专家决策最优的假设进行设计.

然而,由于在函数空间中可能存在多个函数能够同时满足专家策略最优的假设,例如每一步决策所带来的反馈始终为 0 的情况.因此,算法设计的模型应能够解决反馈信号的模糊性(ambiguity).目前,我们可以通过 3 类反馈信号函数的形式实现反馈信号求解过程,它们分别是:1)基于大间隔(max-margin)的反馈信号;2)基于确定基函数组合的反馈信号函数;3)基于参数化的反馈信号函数,例如神经网络.

2.1 基于确定基函数组合的反馈信号函数

逆强化学习发展初期,大多工作均建立在环境反馈信号函数为确定基函数组合的情况下.该类方法通过状态特征构建基函数,从而将求解反馈信号函数的任务转化为求解各个基函数权重的任务.其能够较好地克服反馈信号搜索过程中存在的函数歧义性的问题.

为了建立合适的优化模型求解相关决策问题的反馈信号,该类方法从专家决策轨迹最优的假设出发,通过以下 2 种方法建立相关模型:

1) 根据 1.1 节中的 Bellman 最优等式,我们可以推知:最优策略相对于其他任何策略,在状态空间上具有最大动作价值总和,即对于在动作集中任意非 a_1 的动作都有:$\sum_{s\in S}(Q(s,a_1)-\max_{a\in A\backslash a_1}Q(s,a))$ 的值最大,其中为了便于表示,我们假设 a_1 表示为最优决策动作.因此,逆强化学习的目标也就是使得专

家决策数据所确定的策略具有最高 Q 总和.类似的目标函数还包括:$\sum_{s\in S}\sum_{a\in A}(Q^{\pi}(s,a_1)-Q^{\pi}(s,a))$等[25].

在上述优化目标的基础上,我们可以考虑逆强化学习问题的约束条件还应包括:a_1 为最优决策动作,根据定理 2 可以得知,该条件等价于 a_1 动作在相应状态下的 Q 值将大于其余动作的 Q 值.此外,约束条件中还应保证立即反馈信号值始终是有限值.当考虑到对模型进行正则化时,我们可以得到 Ng 等人[25]提出的针对专家决策轨迹的优化模型,如式(8)所示:

$$\begin{aligned}&\text{maximize}\quad \sum_{i=1}^{N}\min_{a\in\{a_2,\cdots,a_k\}}\{(\boldsymbol{P}_{a_1}(i)-\boldsymbol{P}_a(i))\times\\&(I-\gamma\cdot\boldsymbol{P}_{a_1})^{-1}\}\boldsymbol{R}-\lambda\|\boldsymbol{R}\|_1,\\&\text{s.t.}\ (\boldsymbol{P}_{a_1}(i)-\boldsymbol{P}_a(i))(I-\gamma\cdot\boldsymbol{P}_{a_1})^{-1}\boldsymbol{R}\succ 0,\\&\quad |R_i|\leqslant R_{\max},i=1,2,\cdots,N.\end{aligned}\tag{8}$$

其中,$\boldsymbol{P}_a(i)$表示状态转换概率矩阵.矩阵 $\boldsymbol{R}$ 表示反馈量矩阵.其中模型约束条件

$$(\boldsymbol{P}_{a_1}(i)-\boldsymbol{P}_a(i))(I-\gamma\cdot\boldsymbol{P}_{a_1})^{-1}\boldsymbol{R}\succ 0,$$

表示左侧矩阵各项元素均大于 0,以保证 a_1 为最优决策.$|R_i|\leqslant R_{\max}$亦表示矩阵中各项元素均小于某个有限值.当考虑到决策问题的反馈函数可以由一组确定的基函数线性拟合时,该优化模型可以很好地通过线性规划(linear programming)求解得到相应环境的反馈函数.

2) 根据强化学习基于动态规划算法最大化未来反馈量的经典研究我们可以得知:最优策略相对于其他策略而言将获得最大的未来奖赏,即:

$$V^{\pi}=\sum_{t=0}^{N}R_t(s,a)$$

将取得最大值.

当决策问题的反馈信号可以由一系列确定的基函数 $\varphi_1,\varphi_2,\cdots,\varphi_k$ 线性组合而成时,我们可以定义策略的特征期望[18]为

$$\boldsymbol{\mu}(\pi)=E[\sum_{t=0}^{\infty}\gamma^t\boldsymbol{w}\boldsymbol{\varphi}(s_t)\,|\,\pi]\in R^k.$$

由此,我们可以得到对于任意策略特征期望 μ,可以得到:

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_{\mathrm{E}}\geqslant\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}.$$

其中,$\boldsymbol{\mu}_{\mathrm{E}}$ 表示专家决策数据所确定的专家策略特征期望,其值可以通过蒙特卡洛算法进行估算:

$$\hat{\boldsymbol{\mu}}_{\mathrm{E}}=\frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{\infty}\gamma^t\boldsymbol{\phi}(s_t^{(i)}).$$

通过建立优化模型:

$$\begin{aligned}&\max_{t,\boldsymbol{w}}\ t\\&\text{s.t.}\ \ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_{\mathrm{E}}\geqslant\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}(j)+t,j=0,\cdots,i-1,\\&\quad\|\boldsymbol{w}\|_2\leqslant 1.\end{aligned}\tag{9}$$

我们可以得到以下结论:当优化变量 t 不大于拟合误差 ε 时,算法将得到决策问题反馈信号优化变量 $\boldsymbol{w}$,也即得到未来总反馈函数 $R=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}$.此时,由于 $t\leqslant\varepsilon$,也将得到相应策略,其未来奖赏值$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}(i)\geqslant\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_{\mathrm{E}}-\varepsilon$,也即结合不同正向强化学习策略搜索方法设计的示教学习方法得到的策略将不低于专家策略减去某小量的水平.

通过上述 2 种方式建立的逆强化学习优化模型可以帮助求解得到相关问题的反馈信号.该类方法通过比较专家策略和其他策略的价值,从而建立逆强化学习优化模型,能够较好地实现对专家决策轨迹的学习,并获取环境反馈信号函数.

但是由于在实际结合正向强化学习设计的大多示教学习算法中,逆强化学习模型将被嵌套在策略搜索的循环中,因此降低逆强化学习算法的复杂度得到极大的关注.其中较为经典的工作有 2008 年 Syed 等人[26]提出的 LPAL 算法,通过设计学徒学习的对偶模型,并通过线性规划方法求解该线性模型的占有度(occupancy measure),其算法最坏时间复杂度仅为 $O(|S\|A|+k)$,相对于此前 2004 年 Abbeel 等人[18]提出的学徒学习方法以及 2007 年通过自适应博弈理论优化后的 MWAL 算法[22]在时间效率上都有了极大的提高.

2.2 基于参数化模型的反馈信号函数

随着逆强化学习面对的决策问题复杂度的提升,研究人员开始关注于提升反馈信号函数的表达能力.其中较为有效的是通过参数化模型对环境反馈信号进行建模.

早期的致力于扩大决策问题反馈信号表达能力的工作包括:2010 年 Levine 等人[27]提出的 FIRL (feature construction for IRL)算法,其方法通过构建一组基于逻辑联结的合成特征,从而间接实现了非线性反馈信号的建模.2011 年,Levine 等人[28]又提出了 GP-IRL,其方法采用了基于高斯过程[29]的反馈信号,通过高斯过程极大地增强了反馈函数的表示能力.2015 年,Jin 等人[30]又在 GP-IRL 算法基础上结合了深度信念网络,实现了深度高斯过程在逆强化学习上的应用(DGP-IRL).其中 GP-IRL 和 DGP-IRL 在众多开源环境测试,例如经典的 Grid-world 测试实验以及 gym 下的强化学习基准测试实验中都取得了"state-of-the-art"的效果.

随着深度学习的蓬勃发展,通过神经网络对反馈函数进行建模逐渐称为逆强化学习的一大主流方向.

其中较为知名的是2008年,Ziebart等人[11]提出的最大熵逆强化学习方法(maximum entropy IRL),通过优化专家决策数据集的似然函数实现反馈信号的优化,很好地解决了专家决策数据中可能存在的噪声以及专家数据本身并不是最优的问题.

最大熵逆强化学习方法是经典的基于"能量"的模型(energy-based model)[31].其中能量函数 ε 为环境的代价函数(即反馈信号函数的相反数).根据"能量"模型的假设,可以知道专家在策略轨迹空间的采样概率密度为

$$p(\tau)=\frac{\exp(-c_\theta(\tau))}{\sum_\tau \exp(-c_\theta(\tau))}, \tag{10}$$

其中,τ 为策略轨迹,分母为划分函数 Z(partition function).式(10)可以简单地理解为:当2条决策轨迹具有相同的反馈奖赏时,其具有相同的概率别"专家"采样获得,而当某条轨迹具有更高的反馈奖赏时,"专家"将更有机会能够采样到这条轨迹.

为了让专家决策数据(训练数据)更能够被"专家"采样到,逆强化学习的优化目标是最大化专家轨迹的似然函数,可以表述为

$$\text{maximize} \quad \log \prod_{i=1}^{m} p(\tau_i) = \sum_{i=1}^{m} \log p(\tau_i) = \sum_{i=1}^{m} -\log Z - c_\theta(\tau_i). \tag{11}$$

因此,通过随机梯度方法优化模型式(11)就可以求解得到环境的反馈信号函数.此处需要注意的是:当我们面对的是离散且规模较小的状态空间时,划分函数 Z 可以通过动态规划算法求得;而当我们面对大规模状态空间时,则需要通过采样等方法实现[19,32].

最大熵逆强化学习方法通过优化专家决策数据的似然函数从而获得环境反馈信号,该方法引入了一定的随机性,可以处理专家决策数据本身不是最优或含有一定噪声的情况.

类似能够处理专家决策数据本身不是最优的方法还包括一系列概率模型.其中包括:2007年,Ramachandran和Amir[33]提出贝叶斯非参数化方法去构建反馈函数特征来实现逆强化学习,该方法称作贝叶斯逆强化学习(Bayesian IRL).其后2013年,Choi等人[34]通过构建了一组合成特征上的先验概率优化了该算法.

2.3 其他函数表示形式

对于某些复杂决策问题,环境反馈信号难以通过单一的一个函数进行表示,也就是说是通过单一函数拟合过程中会出现决策数据和反馈函数严重不一致的情况.通过基于每条专家决策轨迹都能够被多个局部一致的反馈函数所生成的假设,Nguyen等人[35]提出了通过期望最大化(expectation-maximization, EM)方法来学习不同的反馈信号以及它们之间动态的转换过程.通过该方法,可以实现针对专家决策轨迹的分割,使得各个部分(segments)均能对应合适的局部一致的反馈函数.基准数据测试(Grid-world以及gym等开源强化学习环境测试)表明该方法也取得了"state-of-the-art"的效果.

此外,逆强化学习领域仍有很多问题需要进行研究解决.例如,在考虑到部分可观察的环境(partially observable environments)[36]时,如何有效地将逆强化学习或示教学习方法迁移到这样的环境之中、如何设计实验来提高反馈函数的可识别性(identifiablity)等问题.

3 基于逆强化学习的示教学习方法

示教学习的目标是通过专家决策轨迹去模仿专家的决策行为.本文第2节介绍了逆强化学习的方法和所需解决的问题,逆强化学习是通过学习专家决策轨迹从而获得环境反馈信号的一类方法.本节将介绍通过结合逆强化学习、正向强化学习策略搜索算法所设计的示教学习方法,也即基于逆强化学习的示教学习方法.

目前,基于逆强化学习的示教学习主要的2个框架分别是:1)在经典的正向强化学习算法内循环中使用逆强化学习算法优化问题的反馈信号,基于反馈信号函数继续实现策略的优化,不断迭代实现示教学习过程.其核心在于将逆强化学习方法置于正向策略搜索方法的内循环之中,经典的方法包括学徒学习方法等.2)基于不断优化得到的反馈信号去实现正向强化学习过程,通过采样数据和专家数据相结合实现逆强化学习过程,同时将正向强化学习过程置于逆强化学习的内循环中,经典的方法有代价指导学习等.本节将主要介绍学徒学习方法和代价指导学习方法.

3.1 学徒学习

学徒学习方法是通过在马尔可夫决策过程中,模仿专家行为,最终得到不差于专家行为策略的方法.其核心的思想是通过匹配专家期望特征实现模仿学习过程.

在线性假设下,反馈信号可以由一组确定基函数 $\varphi_1,\varphi_2,\cdots,\varphi_k$ 进行线性组合.因此,策略的价值可以表示为

$$E_{s_0\sim D}[V^{\pi}(s_0)]=E[\sum_{t=0}^{\infty}\gamma^t R(s_t)\,|\,\pi]=$$

$$E[\sum_{t=0}^{\infty}\gamma^t \boldsymbol{w}\boldsymbol{\phi}(s_t)\,|\,\pi]=\boldsymbol{w}\cdot E[\sum_{t=0}^{\infty}\gamma^t\boldsymbol{\phi}(s_t)\,|\,\pi]. \quad (12)$$

我们可以发现如果有这样一个策略 π,其特征期望满足 $\|\boldsymbol{\mu}(\pi)-\boldsymbol{\mu}_{\mathrm{E}}\|_2\leqslant\varepsilon$ 时,有:

$$\left|E[\sum_{t=0}^{\infty}\gamma^t R(s_t)\,|\,\pi_{\mathrm{E}}]-E[\sum_{t=0}^{\infty}\gamma^t R(s_t)\,|\,\pi]\right|=$$
$$|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}(\pi)-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_{\mathrm{E}}|\leqslant$$
$$\|\boldsymbol{w}\|_2\|\boldsymbol{\mu}(\pi)-\boldsymbol{\mu}_{\mathrm{E}}\|_2\leqslant\varepsilon. \quad (13)$$

因此,我们可以得到以下结论:对于某个策略 π,若其特征期望接近专家策略特征期望,则该策略是学徒学习的一个解.算法 1 描述了由 Abbeel 等人[18]提出的通过结合策略迭代和式(9)的逆强化学习算法所设计的学徒学习方式.

算法 1. 学徒学习算法.

输入:专家决策行为数据;

输出:算法得到的策略以及相应的反馈函数.

① 随机初始化一个策略,计算其特征期望:

$\boldsymbol{\mu}(0)=\boldsymbol{\mu}(\pi(0))$,设置 $i=1$;

② 计算:

$t(i)=\max\limits_{\boldsymbol{w}:\|\boldsymbol{w}\|_2\leqslant 1}\min\limits_{j\in\{0,1,\cdots,(i-1)\}}\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{\mu}_{\mathrm{E}}-\boldsymbol{\mu}(j))$,

并且获得相应 $\boldsymbol{w}$ 值为 $\boldsymbol{w}(i)$;

③ IF $t(i)\leqslant\varepsilon$ THEN

④ 算法终止;

End If

⑤ 使用强化学习算法,计算最优策略 $\pi(i)$ 未来累计奖赏为 $R=(\boldsymbol{w}(i))^{\mathrm{T}}\boldsymbol{\varphi}$;

⑥ 计算策略特征期望 $\boldsymbol{\mu}(i)=\boldsymbol{\mu}(\pi(i))$;

⑦ 设置 $i=i+1$,并返回步骤②.

算法 1 中步骤①初始化随机策略并计算其相应的特征期望;步骤②即为逆强化学习算法(式(9)),其中 $\|\boldsymbol{w}\|$ 表示 $\boldsymbol{w}$ 的二阶范数;步骤③④判断专家策略和目前学得策略的最大策略值的差值,若满足要求则算法终止;步骤⑤基于得到的反馈信号通过正向强化学习算法实现策略优化;步骤⑥⑦计算新策略的特征期望,并迭代.学徒学习方法通过结合逆强化学习(模型 2)和正向强化学习方法实现最优策略的搜索,算法将得到一组策略,通过以下二次规划算法我们可以相应用于决策的混合策略:

$$\min\|\boldsymbol{\mu}_{\mathrm{E}}-\boldsymbol{\mu}\|_2$$
$$\text{s.t. }\boldsymbol{\mu}=\sum_i\lambda_i\boldsymbol{\mu}(i),\ \lambda_i\geqslant 0,$$
$$\sum_{i=1}^{n}\lambda_i=1.$$

其中,λ_i 可以看作是以 λ_i 的概率选择 $\boldsymbol{\mu}(i)$ 策略.

为了将学徒学习方法应用到高性能机器人系统之中,Abbeel 等人[37]通过将学徒学习和探索策略(exploration policies)方法结合解决动态未知的机器人环境.其后,该项工作也被应用到了著名的 Stanford 自动直升机之中.

3.2 代价指导学习

代价指导学习[19]是通过结合正向强化学习中的策略优化[38](policy optimization)方法和最大熵逆强化学习方法[11]实现的示教学习方法.

如图 2 所示,系统通过初始化策略在机器人或设备上进行轨迹采样,并将采样得到的轨迹和专家决策数据进行合并,共同用于实现逆强化学习过程,优化反馈信号函数.基于得到的反馈信号函数,在内循环中实现策略优化.不断迭代上述过程,最终实现示教学习过程.其实现过程如算法 2 所示.

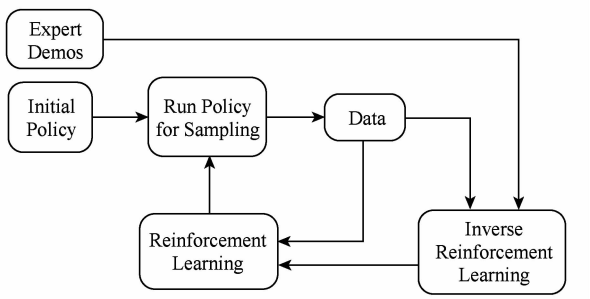

Fig. 2 Guided cost learning procedure

图 2 代价指导学习过程

算法 2. 代价指导学习算法.

输入:专家决策行为数据;

输出:算法得到的策略以及相应的反馈函数.

① 随机初始化一个策略;

② FOR $i=1$ to I DO

③ 通过目前策略采样生成采样数据集;

④ 扩展数据样本集:$D_{\mathrm{samp}}=D_{\mathrm{samp}}\cup D_{\mathrm{traj}}$;

⑤ 通过 D_{samp} 优化问题反馈信号函数;

⑥ 通过正向强化学习更新策略;

⑦ END FOR

⑧ 返回优化后的策略和相应反馈信号.

算法 2 步骤①实现随机初始化策略;步骤③通过当前策略进行采样;步骤④实现采样数据集和专家决策数据集的合并;步骤⑤实现逆强化学习过程(最大熵算法);步骤⑥实现策略优化;不断迭代步骤③~⑥,实现示教学习过程.

此外,由于强化学习系统采样的样本有限,一般可以通过将专家决策数据集合进行分组,通过多组数据循环优化反馈信号,实现逆强化学习过程.目前,

代价指导学习算法在机器人多项智能操作,例如倒水、叠盘子等实验[19]中取得了"state-of-the-art"的效果.

通过上述介绍的学徒学习方法和代价指导学习方法两大类实现框架,我们能够将不同的逆强化学习和正向强化学习方法进行结合从而设计一系列示教学习算法.通过采用不同的逆强化学习方法,我们可以处理专家决策数据本身存在的各种问题,例如数据存在噪声、其决策过程本身并不是最优的以及反馈信号表示能力受到限制等.同样地,通过采用不同的正向强化学习方法,我们可以解决许多由环境所带来的问题,例如实现在高维连续系统中的示教学习应用等.

4 结束语

本文不仅介绍了建立逆强化学习优化模型的方法以及逆强化学习方法发展回顾,还介绍了如何通过结合逆强化学习、正向强化学习方法设计新的示教学习方法,重点介绍了2种框架以及其具有代表性2种的经典方法:学徒学习以及代价指导学习方法.

示教学习是通过模仿专家行为实现专家决策的学习方法.而其中,基于逆强化学习的示教学习方法不仅能够实现针对决策数据的学习,还能够较好地学习到专家行为的动机.

目前,示教学习的主要应用领域为智能机器人操控.在应用过程中,目前示教学习方法也遇到了很多问题,这包括:如何将示教学习算法在不同机器人之间进行迁移[39];如何采用更少量的专家决策数据来学得较好的反馈信号[40]等.此外,将示教学习方法应用到更多的强化学习场景当中也是我们未来的研究方向之一.

参 考 文 献

[1] Sutton R, Barto A. Reinforcement Learning: An Introduction [M]. Cambridge, MA: MIT Press, 2017

[2] Bagnell J. An invitation to imitation [OL]. 2015 [2017-02-12]. https://pdfs.semanticscholar.org/f04d/3ddee335927186b012a1bee765c142ddce57.pdf?_ga=2.181329124.233418181.1525172550-397727813.1525172550

[3] Abbeel P, Ganapathi V, Ng A. Learning vehicular dynamics, with application to modeling helicopters [C] //Proc of the 20th Conf on Neural Information Processing Systems (NIPS). Cambridge, MA: MIT Press, 2006: 1-8

[4] Abbeel P, Coates A, Quigley M, et al. An application of reinforcement learning to aerobatic helicopter flight [C] //Proc of the 21st Conf on Neural Information Processing Systems (NIPS). Cambridge, MA: MIT Press, 2007: 1-8

[5] Ng A, Coates A, Diel M, et al. Inverted autonomous helicopter flight via reinforcement learning [C] //Proc of the 4th Int Symp on Experimental Robotics. New York: Springer, 2004: 799-806

[6] Coates A, Abbeel P, Ng A. Learning for control from multiple demonstrations [C] //Proc of the 25th Int Conf on Machine Learning (ICML). New York: ACM, 2008: 144-151

[7] Abbeel P, Coates A, Hunter T, et al. Autonomous autorotation of an RC helicopter [C] //Proc of the 8th Int Symp on Robotics. New York: Springer, 2008: 385-394

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

[9] Ratliff N, Bagnell J, Zinkevich M. Maximum margin planning [C] //Proc of the 23rd Int Conf on Machine Learning (ICML). New York: ACM, 2006: 729-736

[10] Abbeel P, Dolov D, Ng A, et al. Apprenticeship learning for motion planning with application to parking lot navigation [C] //Proc of the 21st IEEE/RSJ Int Conf on Intelligent Robots and Systems. Piscataway, NJ: IEEE, 2008: 1083-1090

[11] Ziebart B, Maas A, Bagnell J, et al. Maximum entropy inverse reinforcement learning [C] //Proc of the 23rd AAAI Conf on Artificial Intelligence. Menlo Park: AAAI Press, 2008: 1433-1438

[12] Ziebart B, Bagnell J, Dey A. Modeling interaction via the principle of maximum causal entropy [C] //Proc of the 27th Int Conf on Machine Learning (ICML). New York: ACM, 2010: 1255-1262

[13] Barrett E, Linder S. Autonomous HVAC control, a reinforcement learning approach [C] //Proc of the 25th European Conf on Machine Learning and Knowledge Discovery in Databases. Porto: CEUR-WS, 2015: 3-19

[14] Sammut C, Hurst S, Kedzier D, et al. Learning to fly [C] //Proc of the 9th Int Conf on Machine Learning (ICML). New York: ACM, 1992: 385-393

[15] Kunigoshi Y, Inaba M, Inoue H. Learning by watching: Extracting reusable task knowledge from visual observation of human performance [J]. IEEE Transactions on Robotics and Automation, 1994, 10(6): 799-822

[16] Ross S, Bagnell D. Efficient reductions for imitation learning [C] //Proc of the 13th Int Conf on Artificial Intelligence and Statistics (AISTATS). Cambridge, MA: MIT Press, 2010: 661-668

[17] Ross S, Gordon G, Bagnell J. A reduction of imitation learning and structured prediction to no-regret online learning [C] //Proc of the 14th Int Conf on Artificial Intelligence and Statistics (AISTATS). Cadiz: JMLR, 2011: 627-635

[18] Abbeel P, Ng A. Apprenticeship learning via inverse reinforcement learning [C] //Proc of the 21st Int Conf on Machine Learning (ICML). Cadiz: JMLR, 2004: 1-8

[19] Finn C, Levine S, Abbeel P. Guided cost learning: Deep inverse optimal control via policy optimization [C] //Proc of the 33rd Int Conf on Machine Learning (ICML). New York: ACM, 2016: 49-58

[20] Ho J, Ermon S. Generative adversarial imitation learning [C] //Proc of the 30th Conf on Neural Information Processing Systems (NIPS). Cambridge, MA: MIT Press, 2016: 4565-4573

[21] Freund Y, Schapire R. Adaptive game playing using multiplicative weights [J]. Games and Economic Behavior, 1999, 29 (1/2): 79-103

[22] Syed U, Schapire R. A game-theoretic approach to apprenticeship learning [C] //Proc of the 22nd Conf on Neural Information Processing Systems (NIPS). Cambridge, MA: MIT Press, 2008: 1449-1456

[23] Goodfellow I, Abdaie J, Mirza M, et al. Generative adversarial nets [C] //Proc of the 28th Conf on Neural Information Processing Systems (NIPS). Cambridge, MA: MIT Press, 2014: 2672-2680

[24] Finn C, Christiano P, Abbeel P, et al. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models [OL]. (2016-11-11) [2016-11-25]. https://arxiv.org/abs/1611.03852, 2016

[25] Ng A, Russell S. Algorithms for inverse reinforcement learning [C] //Proc of the 17th Int Conf on Machine Learning (ICML). Cambridge, MA: MIT Press, 2000: 663-670

[26] Syed U, Bowling M, Schapire R. Apprenticeship learning using linear programming [C] //Proc of the 25th Int Conf on Machine Learning (ICML). New York: ACM, 2008: 1032-1039

[27] Levine S, Popvic Z. Feature construction for inverse reinforcement learning [C] //Proc of the 24th Conf on Neural Information Processing Systems (NIPS). Cambridge, MA: MIT Press, 2010: 1342-1350

[28] Levine S, Popvic Z. Nonlinear inverse reinforcement learning with gaussian processes [C] //Proc of the 25th Conf on Neural Information Processing Systems (NIPS). Cambridge, MA: MIT Press, 2011: 19-27

[29] Rasmussen C, Williams C. Guassian Processes for Machine Learning. Cambridge, MA: MIT Press, 2006

[30] Jin M, Spanos C. Inverse reinforcement learning via deep gaussian process [OL]. (2015-12-26)[2017-03-04]. https://arxiv.org/abs/1512.08065

[31] LuCun Y, Chopra S, Hadsell R. A tutorial on energy-based learning [J]. Predicting Structured Data, 2006, 1(1): 1-59

[32] Wulfmeier M, Ondruska P, Posner I. Maximum entropy deep inverse reinforcement learning [OL]. (2015-07-17) [2016-03-11]. https://arxiv.org/abs/1507.04888

[33] Ramachandran D, Amir E. Bayesian inverse reinforcement learning [C] //Proc of the 20th Int Joint Conf on Artificial Intelligence (IJCAI). Burlington: Morgan Kaufmann, 2007: 2586-2591

[34] Choi J, Kim K. Bayesain nonparametric feature construction for inverse reinforcement learning [C] //Proc of the 27th Int Joint Conf on Artificial Intelligence (IJCAI). San Francisco, MA: Morgan Kaufmann, 2013: 1287-1293

[35] Nguyen Q, Low K, Jaillet P. Inverse reinforcement learning with locally consistent reward functions [C] //Proc of the 29th Conf on Advances in Neural Information Processing Systems (NIPS). Cambridge, MA: MIT Press, 2015: 1747-1755

[36] Choi J, Kim K. Inverse reinforcement learning in partially observable environments [J]. Journal of Machine Learning Research, 2011, 12(2): 1028-1033

[37] Abbeel P, Ng A. Exploration and apprenticeship learning in reinforcement learning [C] //Proc of the 22nd Int Conf on Machine Learning (ICML). Cambridge, MA: MIT Press, 2005: 1-8

[38] Levine S, Abbeel P. Learning neural network policies with guided policy search under unknown dynamics [C] //Proc of the 28th Conf on Neural Information Processing Systems (NIPS). Cambridge, MA: MIT Press, 2014: 1071-1079

[39] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks [OL]. (2017-03-09) [2017-07-18]. https://arxiv.org/abs/1703.03400

[40] Duan Y, Andrychowicz M, Stadie B, et al. One-shot imitation learning [OL]. (2017-05-21) [2017-12-04]. https://arxiv.org/abs/1703.07326

Zhang Kaifeng, born in 1994. Master candidate of the Department of Computer Science and Technology, Nanjing University. His main research interests include machine learning and optimization.

Yu Yang, born in 1982. PhD and associate professor in the Department of Computer Science and Technology, Nanjing University. His research interests mainly include machine learning and reinforcement learning.