一种不稳定环境下的策略搜索及迁移方法

朱 斐[1,2,3],刘 全[1,3],傅启明[1,3,4],陈冬火[1],王 辉[1],伏玉琛[1]
(1. 苏州大学计算机科学与技术学院,江苏苏州 215006;2. 苏州大学江苏省计算机信息处理技术重点实验室,江苏苏州 215006;
3. 符号计算与知识工程教育部重点实验室(吉林大学),吉林长春 130012;4. 苏州科技学院电子与信息工程学院,江苏苏州 215011)

摘 要: 强化学习是一种 Agent 在与环境交互过程中,通过累计奖赏最大化来寻求最优策略的在线学习方法.由于在不稳定环境中,某一时刻的 MDP 模型在与 Agent 交互之后就发生了变化,导致基于稳定 MDP 模型传统的强化学习方法无法完成不稳定环境下的最优策略求解问题.针对不稳定环境下的策略求解问题,利用 MDP 分布对不稳定环境进行建模,提出一种基于公式集的策略搜索算法——FSPS. FSPS 算法在学习过程中搜集所获得的历史样本信息,并对其进行特征信息的提取,利用这些特征信息来构造不同的用于动作选择的公式,采取策略搜索算法求解最优公式.在此基础之上,给出所求解策略的最优性边界,并从理论上证明了迁移到新 MDP 分布中策略的最优性主要依赖于 MDP 分布之间的距离以及所求解策略在原始 MDP 分布中的性能.最后,将 FSPS 算法用于经典的 Markov Chain 问题,实验结果表明,所求解的策略具有较好的性能.

关键词: 强化学习;策略搜索;策略迁移;不稳定环境;公式集
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112 (2017)02-0257-10
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.02.001

A Policy Search and Transfer Approach in the Non-stationary Environment

ZHU Fei[1,2,3], LIU Quan[1,3], FU Qi-ming[1,3,4], CHEN Dong-huo[1], WANG Hui[1], Fu Yu-chen[1]
(1. *School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu* 215006, *China*;
2. *Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou, Jiangsu* 215006, *China*;
3. *Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin* 130012, *China*;
4. *College of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou, Jiangsu* 215011, *China*)

Abstract: As an online learning algorithm, reinforcement learning, which obtains the optimal policy with the maximum expected cumulative reward by interacting with the environment, is mostly based on the stationary Markov Decision Process (MDP) but however is unable to deal with problems of the non-stationary case because traditional reinforcement learning algorithms cannot be used to learn an optimal policy directly due to the failure of MDP model after the agent once interacts with the environment. Hereby, a novel policy search algorithm based on a formula set (FSPS), which is generated by features extracted from the collected historical sample trajectories, was proposed. The algorithm adopted the formula with the best performance as the optimal policy. The algorithm also took advantage of concept of transfer learning by transferred the learned policy between two similar MDP distributions, where the performance of the transferred policy mainly depends on the distance between two MDP distributions as well as the performance of the learned policy in the original MDP distribution. Simulation results on the Markov Chain problem show that the algorithm can solve the problem of the non-stationary case quite well.

Key words: reinforcement learning; policy search; policy transfer; non-stationary environment; formula set

1 引言

强化学习(Reinforcement Learning, RL)是一种从环境状态到动作映射的学习:强化学习的 Agent 选择动作(action),状态(state)随之发生改变,环境对此给出一个立即奖赏(reward)作为激励信号.强化学习的目标是期望从环境中得到长期最大累计奖赏(return, R)[1,2].基于强化学习的算法通常利用马尔科夫决策过程(Markov

收稿日期:2015-11-03;修回日期:2016-07-14;责任编辑:马兰英
基金项目:国家自然科学基金(No. 61303108, No. 61373094, No. 61272005, No. 61472262, No. 61502329);江苏省高校自然科学研究基金(No. 13KJB520020);吉林大学符号计算与知识工程教育部重点实验室基金(No. 93K172014K04);苏州市应用基础研究计划基金(No. SYG201422);苏州大学高校省级重点实验室基金(No. KJS1524);中国国家留学基金(No. 201606920013)

Decision Process,MDP)进行建模. 在建模过程中,一般假设环境是稳定的,因此在学习过程中所建立的 MDP 模型不会随着时间的变化而改变. 但是,在很多实际情况中,环境虽然在某个较短时间是相对稳定的,从长期来看是并非是稳定而是会发生变化的. 这导致了在原先环境中所建立的 MDP 模型很可能无法适用于新的环境. 进一步分析可以发现,这种不稳定环境可以分解看成由多个生存期较短的"瞬时"稳定环境所组合而成的. 相应的,在这种环境下,强化学习的目标就从获得长期较高累计奖赏转变为获得"瞬时"稳定环境期间较高累计奖赏. 故而,不稳定环境的 Agent 不仅要考虑获取最优策略,还要考虑在学习过程中所搜集的立即奖赏以及其他历史信息,使之在每个"瞬时"稳定环境建立 MDP 模型中,都获得较高的累计奖赏. 但是两个主要的原因致使传统的强化学习算法不能很好地解决此类问题. 首先,由于传统的强化学习算法通常只是求解一个最优策略,不考虑学习过程中所收集到的奖赏值,因此难以实现上述目标. 其次,对于每个"瞬时"MDP,Agent 可以与之交互一次,并取得一个状态样本转移序列,称之为单轨迹样本. 在不稳定环境中,Agent 还必须额外考虑解决单轨迹样本学习过程中平衡探索和利用的难题. 目前虽然也有一些工作将学习过程中的立即奖赏考虑进来[3~6],通过最小化无折扣累计奖赏损失函数加快算法的收敛,如 Micheal 等人通过构造复杂的小公式以引入学习过程中的探索信息来指导 Agent 的动作选择[7],但是这些方法大多都无法求解不稳定环境下的最优策略.

迁移学习是机器学习领域的一个研究热点[8]. 有研究人员利用迁移学习将知识从源任务迁移到目标任务,以提高强化学习的性能[9]. 例如,Sorg 和 Singh 提出利用 MDP 之间的软同态特征,构造不同问题中的状态之间的映射关系,实现不同问题之间的迁移学习,并给出知识迁移过程中损失函数的理论边界[10]; Lazaric 等人在假设目标任务与历史任务具有类似的状态转移函数以及奖赏函数的情况下,将历史任务中收集到的样本数据迁移到目标任务中,并结合目标任务中的样本数据,利用基于批处理的强化学习方法进行学习,以减少目标任务中所需要生成的样本数据,加快算法的收敛[11];Castro 等人通过自模拟度量方法构造不同任务中状态之间的相似性关系,将动作选择策略在相似状态之间进行迁移,并从理论上证明策略迁移的合理性[12,13]. 但是在不稳定环境下,上述的方法通常难以求解一个较优的策略,以达到对于任意"瞬时"MDP 都能够取得较大累计奖赏的目的,因此,无法将上述方法直接用于不稳定环境下策略的求解及迁移.

本文针对不稳定环境下的强化学习问题,利用 MDP 分布来描述不稳定的环境,结合小公式集的构造方法,提出一种基于公式集的策略搜索算法,并从理论上证明策略的最优性边界值. 在此基础之上,利用自模拟度量[13]构造 MDP 分布之间的距离度量公式,将所求解的策略在不同 MDP 分布之间进行迁移,并从理论上证明迁移之后策略的最优性边界.

2 背景知识

MDP 通常可用一个四元组 $<X,U,f,\rho>$ 表示,其中 X 是状态集合,U 是动作集合,f: $X\times U\to X$ 是状态转移函数,ρ: $X\times U\to\mathbb{R}$是奖赏函数. 在时刻 k,状态为 x_k,选择动作 u_k,获得的奖赏值为 r_{k+1},即:

$$r_{k+1}=\rho(x_k,u_k) \tag{1}$$

为了强调 r_{k+1}是在状态 x_k 下所获得的奖赏值,奖赏值也记为 $r(x_k)$,则:$r_{k+1}=r(x_k)=\rho(x_k,u_k)$

如果根据策略 h,从状态 x_k 迁移到下一个状态 x_{k+1},即 $x_{k+1}=f(x_k,h(x_k))$,则奖赏函数可记为

$$r_{k+1}=r(x_k)=\rho(x_k,h(x_k)) \tag{2}$$

对于某状态序列$\{x_0,x_1,\cdots,x_n,x_{n+1}\}$,其中状态 x_i $(0\leqslant i\leqslant n)$的后继状态为 x_{i+1} $(0\leqslant i\leqslant n)$,$x_{n+1}$表示终止状态,所获得的折扣累计奖赏为

$$R^h(x)=\sum_{k=0}^{\infty}\gamma^k r(x_k)=\sum_{k=0}^{\infty}\gamma^k\rho(x_k,h(x_k)) \tag{3}$$

强化学习的算法通常利用动作值函数 $Q^h(x,u)$ 和状态值函数 $V^h(x)$ 对策略 h 进行评估. 动作值函数 $Q^h(x,u)$是指根据策略 h,在状态 x 下,采取动作 u 所获得的累计奖赏,而状态值函数 $V^h(x)$是指根据某个策略 h,从特定状态 x 所获得的累计奖赏[14].

根据定义[14],动作值函数的计算为

$$Q^h(x,u)=\rho(x,u)+\gamma R^h(f(x,u))$$

根据策略 h,采取动作 u,某一个状态 x 的后继状态 $x'=f(x,h(x))=f(x,u)$,因此可以得到

$$Q^h(x,u)=\rho(x,u)+\gamma R^h(f(x,u)) \tag{4}$$

对某状态动作对序列$\{(x_0,u_0),(x_1,u_1),\cdots,(x_n,u_n),(x_{n+1},u_{n+1})\}$,其中状态动作对$(x_i,u_i)$ $(0\leqslant i\leqslant n)$的后继状态动作对为$(x_{i+1},u_{i+1})$ $(0\leqslant i\leqslant n)$,$(x_{n+1},u_{n+1})$表示终止状态动作对,结合式(3)和式(4)可以得到 $Q^h(x_0,u_0)$的折扣累积奖赏为

$$\begin{aligned}Q^h(x_0,u_0)&=\rho(x_0,u_0)+R^h(x_1)\\&=\rho(x_0,u_0)+\sum_{k=1}^{\infty}\gamma^k\rho(x_k,h(x_k))\\&=\rho(x_0,u_0)+\sum_{k=1}^{\infty}\gamma^k\rho(x_k,u_k)\\&=\rho(x_0,u_0)+\gamma\sum_{k=1}^{\infty}\gamma^{k-1}\rho(x_k,u_k)\\&=\rho(x_0,u_0)+\gamma\rho(x_1,u_1)+\gamma\sum_{k=2}^{\infty}\gamma^{k-1}\rho(x_k,u_k)\end{aligned}$$

$$= \rho(x_0,u_0) + \gamma\left[\rho(x_1,u_1) + \sum_{k=2}^{\infty}\gamma^{k-1}\rho(x_k,u_k)\right]$$

$$= \rho(x_0,u_0) + \gamma\left[\rho(x_1,u_1) + \sum_{i=2}^{\infty}\gamma^{i}\rho(x_i,u_i)\right]$$

其中,$\rho(x_1,u_1) + \sum_{i=2}^{\infty}\gamma^{i}\rho(x_i,u_i)$为 $Q^h(x_1,u_1)$的计算方式. 因此可知

$$Q^h(x_0,u_0) = \rho(x_0,u_0) + \gamma Q^h(x_1,u_1)$$

于是可以得到

$$Q^h(x,u) = \rho(x,u) + \gamma Q^h(x',u) \tag{5}$$

其中,x'是 x 的后续状态.

根据定义[14],状态值函数的计算为

$$V^h(x) = R^h(x) \tag{6}$$

类似 $Q^h(x,u)$ 的折扣累积奖赏,可以得到 $V^h(x)$的折扣累积奖赏计算方法

$$V^h(x) = \rho(x) + \gamma V^h(x') \tag{7}$$

其中,x'是 x 的后续状态.

在不稳定环境中,下一个状态是具有随机性的,不能通过当前状态和选择的动作确定,相应的动作值函数 $Q^h(x,u)$和状态值函数 $V^h(x)$也需要考虑不确定性. 于是,在不稳定环境的 MDP 模型中,使用状态转移概率函数$\tilde{f}$: $X\times U\times X\to[0,1]$替代确定环境中的状态转移函数$f$. 这样,根据策略 h,在时刻 k、状态为 x_k、选择动作 u_k,状态转移至后续状态 $x_{k+1}\in X_{k+1}$的概率[14]

$$p(x_{k+1}\in X_{k+1}\mid x_k,u_k) = \int_{X_{k+1}}\tilde{f}(x_k,u_k,x')\,\mathrm{d}x' \tag{8}$$

其中,$X_{k+1}\subseteq X$ 是在时刻 k、状态 x_k 的所有可能后续状态 x_{k+1}的集合,$x_{k+1}\in X_{k+1}$. 与奖赏值的表示类似,如果根据策略 h,从状态 x_k 迁移到下一个状态 x_{k+1},则状态转移概率函数亦可记为 $p^h(x_{k+1}\mid x_k)$.

相应的,在不确定环境下,在状态 x_k 中采取动作 u_k 之后,转移到状态 x_{k+1},其奖赏值也对应的变为

$$\tilde{\rho}(x_k,u) = p(x_{k+1}\mid x_k,u)\,r_{k+1} \tag{9}$$

在不稳定环境中,状态 x 服从分布 $\cdot$,根据策略 h,从某个起始状态 x_0 得到的折扣累计奖赏为

$$R^h(x_0) = E_{x_{k+1}\sim\tilde{f}(x_k,h(x_k),\cdot)}\left\{\sum_{k=0}^{\infty}\gamma^{k}\tilde{\rho}(x_k,h(x_k))\right\} \tag{10}$$

其中:E 表示期望;$h(x_k)$表示根据策略 h,在状态 x_k 下采取的动作;$\tilde{f}$ 是从状态 x_k 迁移到状态 x_{k+1}的状态转移概率函数;$x_{k+1}\sim\tilde{f}(x_k,h(x_k),\cdot)$表示从分布 $\cdot$ 中抽取下一个状态 x_{k+1}.

因此,结合式(5)和式(8),可知不确定环境中的 Q^h 值计算的一般形式为

$$Q^h(x,u) = E_{x'\sim\tilde{f}(x,h(x),\cdot)}\{\tilde{\rho}(x,h(x)) + \gamma Q^h(x',u)\} \tag{11}$$

其中,x'是 x 的后续状态.

不稳定环境还需要考虑奖赏值的不稳定,因此在计算奖赏值的时候再乘以一个概率. 这样,式(11)的等式左部分所表示的不稳定环境的奖赏值的计算相应的变为

$$\tilde{r}(x,u) = \int_{\mathbf{R}}p^h(r\mid x,u)\,r\,\mathrm{d}r \tag{12}$$

其中,$p^h(r\mid x,u)$表示根据策略 h,在状态 x 下,采取动作 u,获得奖赏值 r 的概率.

如果考虑所采取的动作也存在不确定性,即:h 所对应的动作是可采取动作集合中的某一个,那么,在计算 Q^h 的时候也需要考虑采取某个动作的概率,这样,式(11)的等式右部分的计算相应的变为

$$Q^h(x',u) = \int_U p^h(u'\mid x')\,Q^h(x',u')\,\mathrm{d}u' \tag{13}$$

其中,U 是所有可采取的动作集,$p^h(u'\mid x')$是根据策略 h 在状态 x'下采取动作 u'的概率.

因此,可以得到不稳定环境中 Q^h 的计算公式

$$Q^h(x,u) = \int_{\mathbf{R}}p^h(r\mid x,u)\,r\,\mathrm{d}r + \int_U p^h(u'\mid x')\,Q^h(x',u')\,\mathrm{d}u' \tag{14}$$

类似的,可以得到不稳定环境中 V^h 的计算方式

$$V^h(x) = \int_{\mathbf{R}}p^h(r\mid x)\,r\,\mathrm{d}r + \int p^h(x'\mid x)\,V^h(x')\,\mathrm{d}x' \tag{15}$$

在强化学习中,能够获得最大化期望回报的策略称为最优策略 h^*,与之相对应的最优动作值函数为 $Q^*(x,u)$,最优状态值函数为 $V^*(x)$. 因此,对于任意策略 h 及状态动作对(x,u),都存在 $Q^*(x,u)\geqslant Q^h(x,u)$;对于任意策略 h 及状态都存在 $V^*(x)\geqslant V^h(x)$. 虽然一个强化学习问题可能同时存在多个最优策略,但是其最优动作值函数或最优状态值函数却是唯一的,其更新方式如式(16)和式(17)所示

$$Q^*(x,u) = \int_{\mathbf{R}}p^h(r\mid x,u)\,r\,\mathrm{d}r + \gamma\int_U p^h(u'\mid x')\max_{u'}Q^h(x',u')\,\mathrm{d}u' \tag{16}$$

$$V^*(x) = \int_{\mathbf{R}}p^h(r\mid x)\,r\,\mathrm{d}r + \gamma\max\int p^h(x'\mid x)\,V^*(x')\,\mathrm{d}x' \tag{17}$$

接下来,给出有界 MDP 的定义. 本文所讨论的 MDP 都是有界的,也是后续证明所需要满足的前提条件.

定义 1(有界 MDP). 假设 X 和 U 都是一个有限集合,奖赏值函数 ρ 有界,即 $R_{\min}\leqslant\rho(x,u)\leqslant R_{\max}$,其中 $R_{\min}$和 $R_{\max}$是常数;设 $\beta = 1/(1-\gamma)$,其中 $\gamma\in(0,1)$为折扣因子,则在任意策略 h 下,对于 $\forall x\in X$ 及 $\forall(x,u)\in X\times U$,$\beta R_{\min}\leqslant V^h(x)\leqslant\beta R_{\max}$和 $\beta R_{\min}\leqslant Q^h(x,u)\leqslant\beta R_{\max}$成立.

在不稳定环境下,假设 h 为某一策略,$dis^M(\cdot)$为某一 MDP 中的状态的分布(用 M 表示该 MDP),则策略 h 在该“瞬时”MDP 下的性能指标可以表示为

$$J_M^h = E_{x\sim dis^u(\cdot)}\{J_M^h(x)\} \tag{18}$$

其中 $J_M^h(x)$ 的计算方式如式(19)所示

$$J_M^h(x) = \{\sum_{k=0}^{\infty}\gamma^k r(x_k) \mid x_0 = x\} \tag{19}$$

其中 $x_{k+1}\sim p(\cdot\mid x_k, h(x_k))$,$h(x_k)$ 表示根据策略 h 在 x_k 下所采取的动作. 假设所有与 Agent 交互的 MDP 都服从分布 $P_M(\cdot)$,在不稳定环境下,目标从获得某一 MDP 下的最优策略转变为获取某一 MDP 分布 $P_M(\cdot)$ 下的最优策略. 式(20)给出策略在分布 $P_M(\cdot)$ 下的性能指标

$$J_{P_M}^h = E_{M\sim P_M}\{J_M^h\} \tag{20}$$

接下来,讨论 MDP 的分布. 在有模型的强化学习中,我们可以使用模型参数的先验分布来表示. 类似地,我们使用 MDP 分布来表示某一未知 MDP 的不确定性,对分布中的每个 MDP,都可以使用强化学习的方法来解决. 我们将概率分布赋给整个 MDP 集合,这样集合中的每个 MDP 都有一个概率. 不失一般性,我们给出 $P_M(\cdot)$ 的具体形式,并假设任意 MDP 都以某一概率被采样.

假设 1 假设 $P_M(\cdot)$ 满足多项式分布,其中包含 K 个可能的 MDP,$M_1, M_2, \cdots, M_k$,与之相对应的概率分别为 $p_1, p_2, \cdots, p_k$. 因此,在 n 次采样中,假设 M_1 出现 n_1 次,M_2 出现 n_2 次,$\cdots$,M_k 出现 n_k 次,则概率质量函数可以表示为

$$p(M_1 = n_1, \cdots, M_k = n_K) = \frac{n!}{n_1!\ \cdots n_K!}p_1^{n_1}\cdots p_K^{n_k} \tag{21}$$

3 基于公式集的策略搜索

在经典强化学习算法中,常常使用如 ε - 贪心策略(ε-Greedy Policy)、Boltzman 探索策略等软策略(Soft Policy)来解决学习过程中探索和利用的平衡问题. 而事实上,这些策略通常由某一个公式或者函数所决定,如 Boltzman 探索策略是基于动作值函数的. 但是,在不稳定环境中,对于任意"瞬时"MDP,在交互过程中,无法根据有限的样本信息求解最优值函数并有效地平衡探索和利用,以获取较高的长期累计奖赏. 本文根据 Michael 等人提出的小公式集的概念[7],结合交互过程中的历史信息,构造更加复杂的公式,利用复杂公式中所蕴含的启发式信息指导动作选择,以加快求解最优策略的效率.

假设当前时刻为 k,状态为 x_k,轨迹样本为 $H_k = \{x_0, u_0, r_1, x_1, u_1, r_2, x_2, \cdots, x_{k-1}, u_{k-1}, r_k, x_k\}$,所构造的公式可以表示:$F: H\times X\times U\to\mathbb{R}$,则该时刻的动作选择策略可以表示为:$h(H_k, x_k) = \arg\max_{u\in U}F(H_k, x_k, u)$. 令 $\boldsymbol{F}$ 为公式空间,可以通过以下四种方式构造公式 F:

①通过二元操作符构造 $F: F = B(F', F'')$,其中 B 属于二元操作符集合 $\boldsymbol{B}$,$F', F''\in\boldsymbol{F}$.

②通过一元操作符构造 $F: F = U(F')$,其中 U 属于一元操作符集合 $\boldsymbol{U}$,$F'\in\boldsymbol{F}$.

③通过变量构造 $F: F = V$,其中 V 属于某一变量集合 $\boldsymbol{V}$,V 可以是状态值函数、动作值函数、奖赏函数等.

④通过常量构造 $F: F = C$,其中变量 C 属于某一常量集合 $\boldsymbol{C}$.

由于无法直接利用历史信息构造公式,因此,考虑将历史信息提取为一组特征变量集合,$\boldsymbol{V} = \{\rho(x_k, u), N(x_k, u), Q(x_k, u), V(x_k), k, \gamma^k\}$,其中 $N(x_k, u)$ 是状态动作对 (x_k, u) 出现的次数. 另外,定义一组操作符集合 $\boldsymbol{B} = \{+, -, \times, \div, \max, \min\}$ 和 $\boldsymbol{U} = \{\sqrt{(\cdot)}, \ln(\cdot), abs(\cdot)\}$ 以及常数集合 $\boldsymbol{C} = \{1, 3, 5, 7\}$. 根据上述定义,我们可以给出具体的公式,如 $F = abs\left(\frac{\rho(x_k, u)}{N(x_k, u)} + \frac{\ln(Q(x_k, u))}{3}\right)$,并根据公式确定具体的动作选择策略,如式(22)所示

$$h^F(H_k, x_k) = \arg\max_{u\in U}F(\rho(x_k, u), N(x_k, u), Q(x_k, u), V(x_k), k, \gamma^k) \tag{22}$$

其中 h^F 表示根据公式 F 所确定的动作选择策略.

下面给出基于公式集的策略搜索算法(formula set based policy search,FSPS),如算法 1 所示.

算法 1 基于公式集的策略搜索算法——FSPS

输入:MDP 分布 $P_M(\cdot)$ 以及公式集 $\boldsymbol{F}$

输出:带来最大 $\hat{J}_{P_M}^{h_F}$ 值的公式 F,确定策略 h^F

```
初始化:对于任意 F∈F, Ĵ^{h_F}_{P_M} = 0, MDP 样本数目 N, 最大搜索次数 K_max
for i = 1 to |F| do
    采样一组包含 N 个 MDP 的样本集合, MDP_Set = Sample(P_M(·), N)
    将当前公式 F_i 所确定的 h^{F_i} 与每一个 MDP 样本交互, Ĵ^{h_{F_i}}_{P_M} = execute(h^{F_i}, MDP_Set), 其中 Ĵ^{h_{F_i}}_{P_M} 是 J^{h_{F_i}}_{P_M} 的近似值
end for
for i = 1 to K_max do
    采样一组包含 N 个 MDP 的样本集合, MDP_Set = Sample(P_M(·), N)
    初始化 max_F 为一个尽可能小的负值
    for all F∈F do
        得到公式 F 被选择的次数 n_F
        计算 Ĵ^{h_F}_{P_M} + sqrt(2 ln i / n_F)
        if Ĵ^{h_F}_{P_M} + sqrt(2 ln i / n_F) > max_F then
            max_F = Ĵ^{h_F}_{P_M} + sqrt(2 ln i / n_F)
            F_sel = F
        end if
```

16:　end for
17:　$n_{F_{sel}} = n_{F_{sel}} + 1$
18:　将当前公式 F_{sel} 所确定的 $h^{F_{sel}}$ 与每一个 MDP 样本交互,$\hat{J}_{P_M}^{h_F} = execute(h^{F_{sel}}, MDP\_Set)$
19: end for
20: return 带来最大 $\hat{J}_{P_M}{}^{h_F{}_{sel}}$ 值的公式 F_{sel},确定策略 $h^{F_{sel}}$

4　MDP 分布之间的距离

在介绍 MDP 分布之间的距离之前,我们介绍一些预备知识. 首先给出度量的定义,如定义 2 所示.

定义 2（度量）. 定义在状态集合 X 上的一个半度量 $d: X \times X \to [0, \infty)$,对于 $\forall x', x'', x''' \in X$ 满足以下性质:①$x' = x'' \Rightarrow d(x', x'') = 0$;②$d(x', x'') = d(x'', x')$;③$d(x', x''') \leqslant d(x', x'') + d(x'', x''')$. 如果上述性质的逆命题也成立,则被称作状态集 X 上的度量.

Givan 等人将自模拟关系引入 MDP 并度量 MDP 中状态之间关系[15]. 简单而言,如果两个状态之间满足自模拟关系,那么这两个状态应该共享相同的值函数以及最优动作. 接下来,给出自模拟关系的定义,如定义 3 所示.

定义 3（自模拟）. 若关系 $E \subseteq X \times X$ 是一个自模拟关系,则对于任意 $x', x'' \in X, x'Ex''$ 满足以下性质:①对于 $\forall u \in U, \rho(x', u) = \rho(x'', u)$;②对于 $\forall u \in U, \forall C \in X/E, \sum_{t \in C} P_{x'}^{u}(t) = \sum_{t \in C} P_{x''}^{u}(t)$.

其中,X/E 是状态集合 X 关于 E 的等价集合. 若两个状态 $x', x'' \in X$,满足自模拟关系,记作 $x' \sim x''$.

由于自模拟关系太过于严格,对于任意满足自模拟关系的两个状态,只要奖赏函数或者状态转移函数发生微小的变化,都直接导致这两个状态不再满足自模拟关系. 但是实际上这两个状态依然非常类似,也应该具有类似的值函数以及最优动作. Frens 等人在自模拟关系的基础上,利用 Kantorovich 距离,提出一种用于度量两个状态之间距离的自模拟度量[13]. 两个概率分布 P 和 Q 的 Kantorovich 距离 $T_k(d)(P, Q)$ 可以用带约束的线性规划描述为

$$T_k(d)(P,Q) = \max_{c_i, i=1,\cdots,|X|} \sum_{i=1}^{|X|} (P(x_i) - Q(x_i)) c_i$$
$$\text{服从:} \begin{array}{l} \forall i, j.\ c_i - c_j \leqslant d(x_i, x_j) \\ \forall i.\ 0 \leqslant c_i \leqslant 1 \end{array} \tag{23}$$

其中,d 是状态集合 X 上的一个度量.

该线性规划等价于对偶线性规划

$$\min_{l_{ij}, i=1,\cdots,|X|, j=1,\cdots,|X|} \sum_{k,j=1}^{|X|} l_{kj} d(x_k, x_j)$$
$$\text{服从:} \begin{array}{l} \forall k.\ \sum_j l_{kj} = P(x_k) \\ \forall j.\ \sum_k l_{kj} = P(x_j) \\ \forall k, j.\ 0 \leqslant l_{kj} \leqslant 1 \end{array}$$

通过其对偶形式,很容易求解 Kantorovich 距离,其计算的时间复杂度是 $O(|X|^2 \log |X|)$[15].

定理 1　令 D 是定义在状态集 X 上的度量集合,且度量 $d \in D$. 定义 $G: D \to D, G(d)(x', x'') = \max_{u \in U}(C_R d_a(x', x'') + C_P T_K(d)(P_{x'}^{u} - P_{x''}^{u}))$,其中,$d_u(x', x'') = |\bar{r}(x', u) - \bar{r}(x'', u)|$,$\bar{r}(x, u)$ 是状态动作对 (x, u) 下的期望立即奖赏,$C_R + C_P \leqslant 1$,则 G 存在一个最小不动点 $d_\sim$,且 $d_\sim$ 是一个自模拟度量.

Frens 等人的相关研究工作[15]证明了定理 1 的成立. 同时,Frens 等人还证明了,在给定度量误差 ζ 的情况下,可以通过迭代计算逼近最优自模拟度量 $d_\sim$,且需要的迭代次数至少是 $\left\lceil \frac{\ln \zeta}{\ln C_P} \right\rceil$.

接下来,给出两个 MDP 之间距离的度量.

假设 2　两个 MDP,M_1 和 M_2,来自同一个分布 $P_M(\cdot)$,具有相同的状态集合 X 以及动作集合,奖赏函数和状态转移函数不同,即 $M_1 = \langle X, U, f_1, \rho_1 \rangle$,$M_2 = \langle X, U, f_2, \rho_2 \rangle$.

为了区分两个不同 MDP 中的对应状态,分别用 x^{M_1} 和 x^{M_2} 表示.

定义 4　假设两个 MDP,M_1 和 M_2,满足假设 1 的条件,$d_\sim$ 是状态集上的一个自模拟度量. 则两者之间的距离 $d_\sim^M(M_1, M_2)$ 可以表示为 $d_\sim^M(M_1, M_2) = E_{x^{M_1} \sim Dis^{M_1}(\cdot), x^{M_2} \sim Dis^{M_2}(\cdot)} d(x^{M_1}, x^{M_2})$.

根据上述定义,给出计算两个 MDP 之间距离的算法,如算法 2 所示.

算法 2　MDP 之间距离度量

输入:两个 MDP:M_1 和 M_2
输出:M_1 和 M_2 的距离 $d_\sim^M(M_1, M_2)$
1:　初始化:对于任意 $i, j \in [1, |X|]$,$d(x_i^{M1}, x_j^{M2}) = 0$,初始化 C_p, C_R 以及 ζ
2:　for $k = 1$ to $k > \left\lceil \frac{\ln \zeta}{\ln C_P} \right\rceil$ do
3:　　for $i = 1$ to $|X|$ do
4:　　　for $g = 1$ to $|X|$ do
5:　　　　for $j = 1$ to $|U|$ do
6:　　　　　$T_K(d)(f(x_i^{M1}, u_j, \cdot), f(x_g^{M2}, u_j, \cdot))$
7:　　　　end for
8:　　　$d(x_i^{M1}, x_g^{M2}) = \max_{u \in U} \{C_R d_u(x_i^{M1}, x_g^{M2}) + C_P T_K(d)(f(x_i^{M1}, u_j, \cdot), f(x_g^{M2}, u_j, \cdot))\}$
9:　　end for
10:　end for
11: end for
12: return $d_\sim^M(M_1, M_2) = \max_{i=1,\cdots,|X|} d(x_i^{M1}, x_i^{M2})$

在给出两个 MDP 之间距离的基础之上,我们给出

两个 MDP 分布之间距离的定义. 尽管这个距离并不是一个度量,但是在两个 MDP 分布满足假设 1 且其中 MDP 满足假设 2 的条件下,该距离依然可以有效地计算并反映两个 MDP 分布之间的远近关系.

定义 5 假设两个 MDP 分布,P_{M_1} 和 P_{M_2},满足假设 1,且两个分布中的 MDP 满足假设 2,则 P_{M_1} 和 P_{M_2} 之间的距离 $d_{\sim}^{P_M}(P_{M_1},P_{M_2})$ 可以表示为:

$$d_{\sim}^{P_M}(P_{M_1},P_{M_2})=E_{M_1\sim P_M,M_2\sim P_M}\{d_{\sim}^{M}(M_1,M_2)\}=\sum_{M_1\sim P_M,M_2\sim P_M}p(M_1)p(M_2)d_{\sim}^{M}(M_1,M_2) \tag{24}$$

其中 $p(M_1)$,$p(M_2)$ 分别是 M_1 和 M_2 在分布 P_{M_1} 和 P_{M_2} 中的概率.

5 策略最优性边界

策略最优性边界主要衡量所求解策略与最优策略之间的性能误差. 在给出策略最优性边界之前,为了后续证明的方便,给出一些符号的描述,如表 1 所示.

表 1 一些符号的描述

符号	描述
F	公式集
h_F^*	基于公式集所求解的最优策略
h^*	策略空间中的最优策略
$P_M(\cdot)$	MDP 分布
$M_1,M_2,\cdots,M_n$	采样于某一 MDP 分布的 n 个 MDP 样本
J_M^h	策略 h 在 M 中的性能指标
$J_{P_M}^h$	策略 h 在 $P_M(\cdot)$ 中的性能指标
$\hat{J}_{P_M}^h$	$J_{P_M}^h$ 的近似值,且 $\hat{J}_{P_M}^h=\frac{1}{n}\sum_{i=1}^{n}J_{M_i}^h$,$J_{P_M}^h=E_{M\sim P_M}\{J_M^h\}$

引理 1 对于任意 $\tau\in(0,1)$ 以及任意策略 h,至少有 $1-\tau$ 的概率使得式 $|\hat{J}_{P_M}^h-J_{P_M}^h|\leqslant\frac{R_{\max}-R_{\min}}{1-\gamma}\sqrt{\ln\frac{2}{\tau}}{2n}$ 成立.

证明 根据定义 1、式(18)以及 Hoeffding 不等式,可得,

$$\begin{aligned}&P(|\hat{J}_{P_M}^h-J_{P_M}^h|\geqslant\varepsilon)\\&=P(|\frac{1}{n}\sum_{i=1}^{n}J_{M_i}^h-J_{P_M}^h|\geqslant\varepsilon)\\&\leqslant 2\exp(\frac{-2\varepsilon^2n^2}{\sum_{i=1}^{n}(R_{\max}/(1-\gamma)-R_{\min}/(1-\gamma))^2})\\&=2\exp(\frac{-2\varepsilon^2n(1-\gamma)^2}{(R_{\max}-R_{\min})^2})\end{aligned}$$

令 $\tau=2\exp(\frac{-2\varepsilon^2n(1-\gamma)^2}{(R_{\max}-R_{\min})^2})$,求解可得 $\varepsilon=\frac{R_{\max}-R_{\min}}{1-\gamma}\sqrt{\frac{\ln(2/\tau)}{2n}}$. 因此,对于任意 $\tau\in(0,1)$,至少有 $1-\tau$ 的概率使得式 $|\hat{J}_M^h-J_M^h|\leqslant\frac{R_{\max}-R_{\min}}{1-\gamma}\sqrt{\frac{\ln(2/\tau)}{2n}}$ 成立.

引理 2 假设 $|\hat{J}_{P_M}^{h^*}-\hat{J}_{P_M}^{h}|=\delta_{P_M}(h^*,h)$,则对于任意 $\tau\in(0,1)$ 以及任意策略 h,至少有 $1-\tau$ 的概率使得式 $|J_{P_M}^{h^*}-\hat{J}_{P_M}^{h}|\leqslant\frac{R_{\max}-R_{\min}}{1-\gamma}\sqrt{\frac{\ln(2/\tau)}{2n}}+\delta_{P_M}(h^*,h)$ 成立.

证明 对 $|J_{P_M}^{h^*}-\hat{J}_{P_M}^{h}|$ 推导可得,

$$\begin{aligned}&|J_{P_M}^{h^*}-\hat{J}_{P_M}^{h}|\\&=|J_{P_M}^{h^*}-\hat{J}_{P_M}^{h^*}+\hat{J}_{P_M}^{h^*}-\hat{J}_{P_M}^{h}|\\&\leqslant|J_{P_M}^{h^*}-\hat{J}_{P_M}^{h^*}|+|\hat{J}_{P_M}^{h^*}-\hat{J}_{P_M}^{h}|\\&=|J_{P_M}^{h^*}-\hat{J}_{P_M}^{h^*}|+\delta_{P_M}(h^*,h)\end{aligned}$$

再根据引理 1,对于任意 $\tau\in(0,1)$,至少有 $1-\tau$ 的概率使 $|J_{P_M}^{h^*}-\hat{J}_{P_M}^{h}|\leqslant\frac{R_{\max}-R_{\min}}{1-\gamma}\sqrt{\frac{\ln(2/\tau)}{2n}}+\delta_{P_M}(h^*,h)$ 成立.

定理 2 对于任意 $\tau\in(0,1)$ 以及任意策略 h,至少有 $1-\tau$ 的概率使得式 $|J_{P_M}^{h^*}-J_{P_M}^{h}|\leqslant 2\frac{R_{\max}-R_{\min}}{1-\gamma}\sqrt{\frac{2\ln(2/\tau)}{n}}+\delta_{P_M}(h^*,h)$ 成立.

证明 对 $|J_{P_M}^{h^*}-J_{P_M}^{h}|$ 推导可得,

$$\begin{aligned}&|J_{P_M}^{h^*}-J_{P_M}^{h}|\\&=|J_{P_M}^{h^*}-\hat{J}_{P_M}^{h}+\hat{J}_{P_M}^{h}-J_{P_M}^{h}|\\&\leqslant|J_{P_M}^{h^*}-\hat{J}_{P_M}^{h}|+|\hat{J}_{P_M}^{h}-J_{P_M}^{h}|\end{aligned}$$

再根据引理 1 以及引理 2 可得,可得对于任意 $\tau\in(0,1)$,至少有 $1-\tau$ 的概率使得 $|J_{P_M}^{h^*}-J_{P_M}^{h}|\leqslant 2\frac{R_{\max}-R_{\min}}{1-\gamma}\sqrt{\frac{n^2/\tau}{2\ln n}}+\delta_{P_M}(h^*,h)$ 成立.

引理 3 假设 $d_{\sim}^{P_M}(P_{M_1},P_{M_2})=\theta$,$C_P\leqslant\frac{1}{1+\gamma}$,$C_R\leqslant\frac{C_P}{\gamma}$,则 $|J_{P_{M_1}}^{h_F^*}-J_{P_{M_2}}^{h_F^*}|\leqslant\frac{\theta}{C_R}$,$|J_{P_{M_1}}^{h_{F'}^*}-J_{P_{M_2}}^{h_{F'}^*}|\leqslant\frac{\theta}{C_R}$.

证明 根据式(18)和式(20)可知,

$$\begin{aligned}J_{P_{M_1}}^{h_F^*}&=E_{M_1\sim P_{M_1}}\{J_{M_1}^{h_F^*}\}\\&=E_{M_1\sim P_{M_1}}E_{x^{M_1}\sim dis^{M_1}(\cdot)}\{J_{M_1}^{h_F^*}(x^{M_1})\}\\J_{P_{M_2}}^{h_F^*}&=E_{M_2\sim P_{M_2}}\{J_{M_2}^{h_F^*}\}\\&=E_{M_2\sim P_{M_2}}E_{x^{M_2}\sim dis^{M_2}(\cdot)}\{J_{M_2}^{h_F^*}(x^{M_2})\}\end{aligned}$$

因此,

$$\begin{aligned}&|J_{P_{M_1}}^{h_F^*}-J_{P_{M_2}}^{h_F^*}|\\&=|E_{M_1\sim P_{M_1}}E_{x^{M_1}\sim dis^{M_1}(\cdot)}\{J_{M_1}^{h_F^*}(x^{M_1})\}\end{aligned}$$

$$-E_{M_2\sim P_{M_2}}E_{x^{M_2}\sim dis^{M_2}(\cdot)}\{J^{h_F^*}_{M_2}(x^{M_2})\}|$$
$$=E_{M_1\sim P_M,M_2\sim P_{M2}}E_{x^{M_1}\sim dis^{M_1}(\cdot),x^{M_2}\sim dis^{M_2}(\cdot)}|J^{h_F^*}_{M_1}(x^{M_1})-J^{h_F^*}_{M_2}(x^{M_2})|$$

由式(19)及 Ferns 等人的研究工作[13],可得,

$$|J^{h_F^*}_{P_{M_1}}-J^{h_F^*}_{P_{M_1}}|$$
$$=E_{M_1\sim P_M,M_2\sim P_{M_1}}E_{x^{M_1}\sim dis^{M_1}(\cdot),x^{M_2}\sim dis^{M_2}(\cdot)}$$
$$|E_{u\sim h_F^*}\{\rho_{M_1}(x^{M_1},u)+\gamma\sum_{x'\in X}f_{M_1}(x^{M_1},u,x')V^{h_F^*}(x')\}$$
$$-E_{u\sim h_F^*}\{\rho_{M_2}(x^{M_2},u)+\gamma\sum_{x'\in X}f_{M_2}(x^{M_2},u,x')V^{h_F^*}(x')\}|\tag{25}$$

很明显,式(25)不大于所有动作集合中计算出的最大值,即:

$$E_{M_1\sim P_M,M_2\sim P_{M_1}}E_{x^{M_1}\sim dis^{M_1}(\cdot),x^{M_2}\sim dis^{M_2}(\cdot)}$$
$$\max_{u\in U}|\{\rho_{M_1}(x^{M_1},u)+\gamma\sum_{x'\in X}f_{M_1}(x^{M_1},u,x')V^{h_F^*}(x')\}$$
$$-\{\rho_{M_2}(x^{M_2},u)+\gamma\sum_{x'\in X}f_{M_2}(x^{M_2},u,x')V^{h_F^*}(x')\}|$$

在式(25)两端各乘以 C_R,可以得到,

$$C_R|J^{h_F^*}_{P_{M_1}}-J^{h_F^*}_{P_{M_1}}|$$
$$\leqslant C_RE_{M_1\sim P_M,M_2\sim P_{M_1}}E_{x^{M_1}\sim dis^{M_1}(\cdot),x^{M_2}\sim dis^{M_2}(\cdot)}$$
$$\max_{u\in U}|\{\rho_{M_1}(x^{M_1},u)+\gamma\sum_{x'\in X}f_{M_1}(x^{M_1},u,x')V^{h_F^*}(x')\}$$
$$-\{\rho_{M_2}(x^{M_2},u)+\gamma\sum_{x'\in X}f_{M_2}(x^{M_2},u,x')V^{h_F^*}(x')\}|$$
$$=E_{M_1\sim P_M,M_2\sim P_{M2}}E_{x^{M_1}\sim dis^{M_1}(\cdot),x^{M_2}\sim dis^{M_2}(\cdot)}$$
$$\max_{u\in U}|C_R\{\rho_{M_1}(x^{M_1},u)+\gamma\sum_{x'\in X}f_{M_1}(x^{M_1},u,x')V^{h_F^*}(x')\}$$
$$-C_R\{\rho_{M_2}(x^{M_2},u)+\gamma\sum_{x'\in X}f_{M_2}(x^{M_2},u,x')V^{h_F^*}(x')\}|$$

整理不等式右边,可以得到

$$C_R|J^{h_F^*}_{P_{M_1}}-J^{h_F^*}_{P_{M_1}}|$$
$$\leqslant E_{M_1\sim P_M,M_2\sim P_{M2}}E_{x^{M_1}\sim dis^{M_1}(\cdot),x^{M_2}\sim dis^{M_2}(\cdot)}$$
$$\max_{u\in U}|C_R\{\rho_{M_1}(x^{M_1},u)-\rho_{M_2}(x^{M_2},u)\}+C_R\{\gamma\sum_{x'\in X}$$
$$f_{M_1}(x^{M_1},u,x')V^{h_F^*}(x')-\gamma\sum_{x'\in X}f_{M_2}(x^{M_2},u,x')V^{h_F^*}(x')\}|$$

由条件 $C_R\leqslant\frac{C_P}{\gamma}$可得到,

$$C_R|J^{h_F^*}_{P_{M_1}}-J^{h_F^*}_{P_{M_1}}|$$
$$\leqslant E_{M_1\sim P_M,M_2\sim P_{M_1}}E_{x^{M_1}\sim dis^{M_1}(\cdot),x^{M_2}\sim dis^{M_2}(\cdot)}$$
$$\max_{u\in U}|C_R\{\rho_{M_1}(x^{M_1},u)-\rho_{M_2}(x^{M_2},u)\}+C_P\{\sum_{x'\in X}$$
$$\cdot f_{M_1}(x^{M_1},u,x')V^{h_F^*}(x')-\sum_{x'\in X}f_{M_2}(x^{M_2},u,x')V^{h_F^*}(x')\}|$$

根据式(23)关于 $T_k(d)(P,Q)$的定义[13]可以得到,

$$\sum_{x'\in X}f_{M_1}(x^{M_1},u,x')V^{h_F^*}(x')=C_PT_K(d_\sim)(f_{M_1}(x^{M_1},u,x'))$$
$$\sum_{x'\in X}f_{M_2}(x^{M_2},u,x')V^{h_F^*}(x')=C_PT_K(d_\sim)(f_{M_2}(x^{M_2},u,x'))$$

将其代入上述不等式,可以得到,

$$C_R|J^{h_F^*}_{P_{M_1}}-J^{h_F^*}_{P_{M_1}}|$$
$$\leqslant E_{M_1\sim P_M,M_2\sim P_{M2}}E_{x^{M_1}\sim dis^{M_1}(\cdot),x^{M_2}\sim dis^{M_2}(\cdot)}$$
$$\max_{u\in U}|C_R\{\rho_{M_1}(x^{M_1},u)-\rho_{M_2}(x^{M_2},u)\}+C_PT_K(d_\sim)$$
$$\cdot((f_{M_1}(x^{M_1},u,x'),f_{M_2}(x^{M_2},u,x'))|$$
$$\leqslant E_{M_1\sim P_M,M_2\sim P_{M2}}E_{x^{M_1}\sim dis^{M_1}(\cdot),x^{M_2}\sim dis^{M_2}(\cdot)}$$
$$\max_{u\in U}|C_R\{\rho_{M_1}(x^{M_1},u)-\rho_{M_2}(x^{M_2},u)\}|+\max_{u\in U}|C_P$$
$$\cdot T_K(d_\sim)((f_{M_1}(x^{M_1},u,x'),f_{M_2}(x^{M_2},u,x'))|$$

由定理 1 可得,

$$d_u(x^{M_1},x^{M_2})=\max_{u\in U}C_R\{\rho_{M_1}(x^{M_1},u)-\rho_{M_2}(x^{M_2},u)\}$$

将其代入上述不等式,可以得到,

$$C_R|J_{P_{M_1}}{}^{h_F^*}-J_{P_{M_1}}{}^{h_F^*}|$$
$$\leqslant E_{M_1\sim P_{M_1},M_2\sim P_M}E_{x^{M_1}\sim dis^{M_1}(\cdot),x^{M_2}\sim dis^{M_2}(\cdot)}$$
$$\max_{u\in U}d_\sim(x^{M_1},x^{M_2})+\max_{u\in U}|C_PT_K(d_\sim)$$
$$\cdot((f_{M_1}(x^{M_1},u,x'),f_{M_2}(x^{M_2},u,x'))|$$

由算法 2 第 8 步可知,

$$d(x^{M_1},x^{M_2})=\max_{u\in U}C_PT_K(d)((f(x^{M_1},u,\cdot),f(x^{M_2},u,\cdot))$$

因此,

$$C_R|J_{P_{M_1}}{}^{h_F^*}-J_{P_{M_1}}{}^{h_F^*}|$$
$$\leqslant E_{M_1\sim P_{M_1},M_2\sim P_M}E_{x^{M_1}\sim dis^{M_1}(\cdot),x^{M_2}\sim dis^{M_2}(\cdot)}d_\sim(x^{M_1},x^{M_2})$$
$$=E_{M_1\sim P_{M_1},M_2\sim P_{M2}}d^M_\sim(M_1,M_2)$$
$$=d_\sim{}^{P_M}(P_{M_1},P_{M_2})$$
$$=\theta$$

故而,对于任意状态 $x\in X$,都有式$|J^{h_F^*}_{P_{M_1}}(x)-J^{h_F^*}_{P_{M_1}}\cdot(x)|\leqslant\frac{\theta}{C_R}$成立.

同理,我们可以扩展到最优策略 h^* 中,即对于任意状态 $x\in X$,$|J^{h^*}_{P_{M_1}}(x)-J^{h^*}_{P_{M_1}}(x)|\leqslant\frac{\theta}{C_R}$.

定理 3　假设 $d^{P_M}_\sim(P_{M_1},P_{M_2})=\theta$,$C_P\leqslant\frac{1}{1+\gamma}$,$C_R\leqslant\frac{C_P}{\gamma}$,$h_F^*$ 是根据算法 1 在 $P_{M_1}(\cdot)$中所求得的最优策略.对于任意 $\tau\in(0,1)$,至少有 $1-\tau$ 的概率使得下式成立,

$$|J^{h^*}_{P_{M_1}}-J^{h_F^*}_{P_{M_1}}|\leqslant\frac{2\theta}{C_R}+2\frac{R_{\max}-R_{\min}}{1-\gamma}\sqrt{\frac{\ln(2/\tau)}{2n}}+\delta_{P_{M_1}}(h^*,h)$$

证明　对$|J^{h^*}_{P_{M_1}}-J^{h_F^*}_{P_{M_1}}|$推导可得,

$$|J^{h^*}_{P_{M_1}}-J^{h_F^*}_{P_{M_1}}|$$
$$=|(J^{h^*}_{P_{M_1}}-J^{h^*}_{P_{M_1}})+(J^{h^*}_{P_{M_1}}-J^{h_F^*}_{P_{M_1}})+(J^{h_F^*}_{P_{M1}}-J^{h_F^*}_{P_{M_1}})|$$
$$\leqslant|J^{h^*}_{P_{M_1}}-J^{h^*}_{P_{M_1}}|+|J^{h^*}_{P_{M_1}}-J^{h_F^*}_{P_{M_1}}|+|J^{h_F^*}_{P_{M1}}-J^{h_F^*}_{P_{M_1}}|$$

再根据定理 2 可得,

$$|J^{h^*}_{P_{M_1}}-J^{h_F^*}_{P_{M_1}}|\leqslant|J^{h^*}_{P_{M_1}}-J^{h^*}_{P_{M_1}}|+|J^{h_F^*}_{P_{M1}}-J^{h_F^*}_{P_{M_1}}|$$
$$+2\frac{R_{\max}-R_{\min}}{1-\gamma}\sqrt{\frac{\ln(2/\tau)}{2n}}+\delta_{P_{M_1}}(h^*,h_F^*)$$

再根据引理 3 可得,可得对于任意 $\tau\in(0,1)$,至少

有 $1-\tau$ 的概率使得 $|J^{h^*}_{P_{M_1}} - J^{h^*_F}_{P_{M_1}}| \leqslant \frac{2\theta}{C_R} + 2\frac{R_{\max}-R_{\min}}{1-\gamma}\sqrt{\frac{2\ln\frac{2}{\tau}}{n}} + \delta_{P_{M_1}}(h^*, h^*_F)$ 成立.

根据定理3,我们可以发现策略 h^*_F 从 $P_{M_1}(\cdot)$ 迁移到 $P_{M_2}(\cdot)$ 后,当 MDP 样本足够大时,其性能主要依赖于 $P_{M_1}(\cdot)$ 和 $P_{M_2}(\cdot)$ 之间的距离 $d^{P_M}_{\sim}(P_{M_1}, P_{M_2})$ 以及 h^*_F 在 $P_{M_1}(\cdot)$ 中的 $\delta_{P_{M_1}}(h^*, h^*_F)$. 因此,如果当两个 MDP 分布足够接近时,我们可以直接将策略从一个 MDP 分布直接迁移到另一个 MDP 分布,实现策略的迁移,并保证所迁移策略在新分布中具有较好的性能. 另外,当我们固定 τ 的值,

且令 $\frac{2\theta}{C_R} + 2\frac{R_{\max}-R_{\min}}{1-\gamma}\sqrt{\frac{\ln 2(2/\tau)}{2n}} + \delta_{P_{M_1}}(h^*, h^*_F) = \sigma$ 时,

通过计算可得

$$n = \frac{2\ln(\frac{2}{\tau})C_R^2(R_{\max}-R_{\min})^2}{(C_R\sigma - C_R\delta_{P_{M_1}}(h^*, h^*_F) - 2\theta)^2(1-\gamma)^2},$$

即至少以 $1-\tau$ 概率在采样 n 次之后能够保证最优策略 h^* 和迁移策略 h^*_F 之间的性能误差在 σ 以内.

6 实验结果分析

马尔可夫链(Markov Chain)是马尔可夫过程中的一个特例. 在马尔可夫链中,系统根据概率分布,可以从一个状态变到另一个状态,也可以保持当前状态,马尔可夫链是具有马尔可夫性质的随机变量的一个数列[16].

本文以 Markov Chain 问题为实验平台验证算法所求解策略的性能以及该策略被迁移后的最优性边界. 在该 Markov Chain 问题实验中,有两个 MDP 分布,P_{M_1} 以及 P_{M_2},一个具有状态 $\{x_1, x_2, x_3, x_4, x_5\}$ 的状态集,以及一个具有动作 $\{u_1, u_2\}$ 的动作集. 在任意状态下,采取任意动作都有一定的概率转移到后续状态或者转移到第一状态,例如,在状态 x_3 下采取动作 u_1 以 80% 的概率转移到状态 x_4,反之转移到状态 x_1. 当转移到状态 x_5 时,奖赏是 91,其他情况下,奖赏是 12. 图 1 是在动作 u_1 下的状态转移示意图.

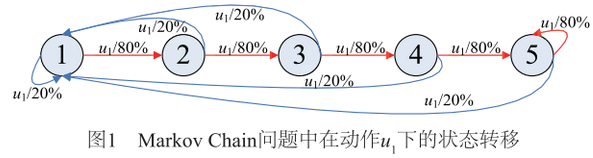

图1 Markov Chain问题中在动作u_1下的状态转移

在实验中,为了方便计算,假定转移概率以及奖赏值分别服从均匀分布,即在区间 $[a,b]$ 上,随机变量 x 的概率密度函数为 $f(x) = 1/(b-a)$, $a \leqslant x \leqslant b$. 状态转移概率按照均匀分布进行采样,分别构造 MDP 分布 P_{M_1} 以及 P_{M_2},其中 P_{M_1} 用于求解最优策略,而 P_{M_2} 用于迁移所学习的最优策略,并测试所迁移策略的性能. 在解决其他问题时,可以根据实际情况设置为其他概率分布,如泊松分布(Poisson distribution)、高斯分布(Gaussian distribution)等.

在实验过程中,设定每个 MDP 分布包含 200 个子 MDP,即 $K=200$;MDP 分布之间距离的阈值 $\zeta = 0.01$,折扣因子 $\gamma = 0.9$, $\tau = 0.01$, $C_R = 0.3$, $C_P = 0.7$, ε-贪心策略中 $\varepsilon = 0.7$;从 MDP 分布中采样得到的子 MDP 的数量默认是 200;根据算法 2,求解 P_{M_1} 和 P_{M_2} 之间的距离是 34.54.

在不稳定 MDP 环境下,包括贪心策略(Greedy Policy)、ε-贪心策略或者模拟退火策略等在内的传统强化学习无法很好地平衡算法执行过程中的探索和利用问题. 例如,贪心策略在算法执行过程中仅利用 Agent 已经学习到的动作,不强调探索,因此有可能会错失更优的动作;ε-贪心策略虽然在利用已经学习到的动作时,会随机进行动作的探索,但是由于这种探索方式是随机的、盲目的,因此无法保证在任意 MDP 下都能得到较优的累计奖赏值. 在稳定 MDP 环境下,算法可以不需要关注学习过程中所收集到的奖赏值,而关注是否能获得最优策略,因此,算法可以利用 ε-贪心策略进行动作的探索;然而,在不稳定环境下,ε-贪心策略的随机探索行为可能带来较小累计奖赏值的动作,这一点是算法应该尽量避免的.

首先,我们通过实验比较各策略在 P_{M_1} 和 P_{M_2} 中的性能. 表 2 给出在 P_{M_1} 和 P_{M_2} 下各策略在执行 50 个时间步后所能带来的累计奖赏值,其中 P_{M_1} 和 P_{M_2} 之间的距离是 34.54. 从表 2 中可以看出,除了最优策略以外,$|3-\sqrt{\rho(x,u)}|$、$|\rho(x,u)-\sqrt{V(x,u)}|$ 以及 $|\rho(x,u)-t|$ 的累计奖赏值较大,说明这些策略能带来较优的执行性能,而贪心策略、贪心策略和随机策略(Random Policy)较差. 从表的右侧数据可以发现,将策略迁移到新的 MDP 分布 P_{M_2} 时,$|3-\sqrt{\rho(x,u)}|$、$|\rho(x,u)-\sqrt{V(x,u)}|$ 以及 $|\rho(x,u)-t|$ 所对应的策略仍然可以获得较大的累计奖赏,说明这些策略迁移后,仍然可以具有较优的性能,而贪心策略、ε-贪心策略和随机策略依然效果不佳.

表 2 P_{M_1} 和 P_{M_2} 中策略性能比较

策略	性能	
	P_{M_1}	P_{M_2}
Optimal Policy	298.28	295.47
$\lvert 3-\sqrt{\rho(x,u)}\rvert$	136.30	131.06
$\lvert\rho(x,u)-\sqrt{V(x,u)}\rvert$	131.62	126.59
$\lvert\rho(x,u)-t\rvert$	125.78	127.06
Greedy Policy	104.13	104.45
ε-Greedy Policy	103.68	102.12
Random Policy	104.20	95.54

接着,我们通过实验比较各策略在不同 MDP 样本数量下的性能. 在实验中,我们设定 MDP 样本的数量分别是 10、20、40、80、150 以及 200. 图 2 是在不同 MDP 采样情况下,与各策略的性能比较图. 从图 2 中可以看出在不同 MDP 样本情况下,相比与贪心策略、ε-贪心策略以及随机策略,$|3-\sqrt{\rho(x,u)}|$ 对应的策略始终能够取得较优的实验结果. 同时,观察 $|3-\sqrt{\rho(x,u)}|$ 在同 MDP 样本情况下累计奖赏值,可以发现 $|3-\sqrt{\rho(x,u)}|$ 的曲线相对是比较稳定的,这也是由于该策略能够较好的平衡学习过程中的探索和利用问题. 另外,相对于基于值函数的动作选择策略,$|3-\sqrt{\rho(x,u)}|$ 更加容易计算,这也可以在一定程度上,加快算法的执行速度.

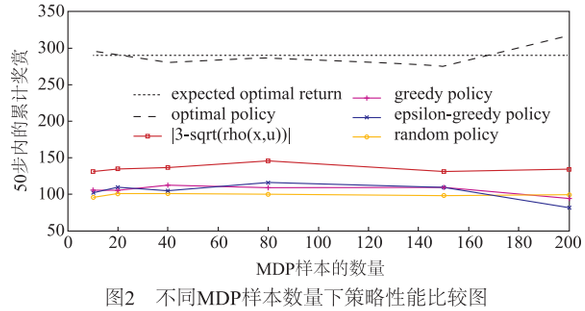

图2　不同MDP样本数量下策略性能比较图

最后,我们实验比较从原始 MDP 分布 P_{M_1} 中所学习到的最优策略迁移到目标 MDP 分布 P_{M_2} 的最优性边界. 在本阶段的实验中,设置不同 MDP 样本数量下迁移策略的最优性边界. 图 3 给出不同 MDP 样本数量下迁移策略的最优性边界,其中上图是迁移策略在 P_{M_2} 下的最优性边界,下图是在不同样本数量情况下,所迁移策略在 P_{M_1} 下累计奖赏值与最优值的近似误差 $\delta_{P_{M_1}}(h^*,h_F^*)$. 从图 3 中可以发现,随着样本数量的增加,所迁移策略在 P_{M_2} 最优性边界将逐渐收敛至一个相对稳定的值,同时,在 P_{M_1} 中被迁移策略的最优性误差在前期震荡之后,也逐渐收敛至一个稳定的值. 另外,在实验过程中,最优性边界大约是 289. 63,策略在 P_{M_1} 的近似误差大约是 159. 因此根据定理 3,我们可以大约计算出,如果需要达到该最优性边界,算法至少需要 2099 个 MDP 样本,而从图 3 中可以看出,在 MDP 的样本数量达到 2100 左右后,最优性边界逐渐趋向于一个稳定的值.

通过实验可以看出,两个具有相同状态空间和动作空间分布的 MDP 之间是可以进行最优策略迁移的,在采样样本数量达到一定规模时,迁移之后策略的最优性边界依赖于两个 MDP 之间的距离以及所迁移策略在原始的 MDP 中的性能. 这与本文相关的理论证明和分析一致. 实验结果表明,本文所提出的 FSPS 算法在解决 Markov Chain 问题时,所求解的策略具有较好的性能.

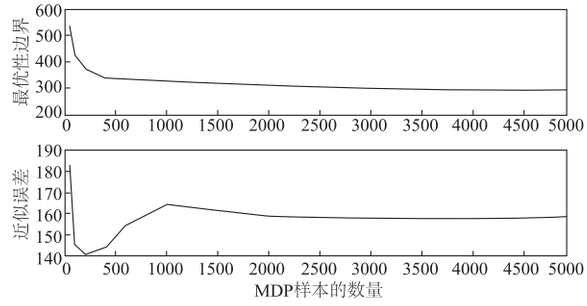

图3　不同MDP样本数量下迁移策略的最优性边界

7　结束语

本文主要针对传统强化学习算法无法求解非稳定 MDP 环境下的预测及控制问题,提出利用 MDP 分布来描述不稳定 MDP 环境,并设计了一种基于公式集的策略搜索算法. 该算法在学习过程中搜集所获得历史样本信息,并对其进行特征信息的提取,利用该特征信息构造不同的用于动作选择的公式,最终利用策略搜索算法求解最优公式,并证明基于该公式的最优策略的最优性边界. 此外,在自模拟度量方法的基础上,本文给出 MDP 距离及 MDP 分布之间距离的定义,并给出 MDP 分布之间距离的计算方法. 结合基于公式集的策略搜索算法所求得的最优策略,将该策略在不同的 MDP 分布之间进行迁移,并证明所迁移的最优策略在新的 MDP 分布中性能的最优性边界主要依赖于两个 MDP 分布之间的距离以及该策略在原始 MDP 分布中的性能.

然而,在计算两个 MDP 分布之间距离时,算法所需要的计算量很大. 在实验中,我们发现当 MDP 分布中 MDP 的数量超过 300 且每个 MDP 中状态的数量超过 50 之后,个人计算机难以计算两个 MDP 分布之间的距离. 本文提出的距离计算方法难以应用于大状态空间和连续状态空间问题中. 因此,下一步的工作就是考虑研究更加高效、简洁的 MDP 分布之间的距离计算方法. 另外,公式集中所构造的公式的优劣直接影响最终算法所获得的最优策略,因此,更加合理地构造公式集也是将来所要继续研究的问题.

参考文献

[1] 朱斐,刘全,傅启明,伏玉琛. 一种用于连续动作空间的最小二乘行动者-评论家方法[J]. 计算机研究与发展,2014,51(3):548-558.
Zhu Fei, Liu Quan, Fu Qiming, Fu Yuchen. A least square actor-critic approach for continuous action space[J]. Journal of Computer Research and Development, 2014, 51(3): 548-558. (in Chinese)

[2] Xu X, Zuo L, Huang Z. Reinforcement learning algorithms with function approximation: recent advances and applica-

tions[J]. Information Sciences,2014,261(5):1 – 31.
[3] 仵博,郑红燕,冯延蓬,等. 一种基于模型的可分解贝叶斯在线强化学习[J]. 电子学报,2014,7(7):1429 – 1434.
Wu Bo,Zheng Hongyan,Feng Yanpeng,et al. Model-based factored bayesian online reinforcement learning[J]. Acta Electronica Sinica,2014,7(7):1429 – 1434. (in Chinese)
[4] 陈学松,刘富春. 一类非线性动态系统基于强化学习的最优控制[J]. 控制与决策,2013,12(12):1889 – 1893.
Chen Xuesong,Liu Fuchun. Optimal control of a class of nonlinear dynamic systems based on reinforcement learning[J]. Control and Decision,2013,12(12):1889 – 1893. (in Chinese)
[5] 赵凤飞,覃征. 一种多动机强化学习框架[J]. 计算机研究与发展,2013,2(2):240 – 247.
Zhao Fengfei, Qin Zheng. A multi-motive reinforcement learning framework[J]. Journal of Computer Research and Development,2013,2(2):240 – 247. (in Chinese)
[6] Jaksch T,Ortner R,Auer P. Near-optimal regret bounds for reinforcement learning[J]. The Journal of Machine Learning Research,2010,11(1):1563 – 1600.
[7] Castronovo M,Maes F,Fonteneau R,et al. Learning exploration/exploitation strategies for single trajectory reinforcement learning[J]. Journal of Machine Learning Research,2012,24:1 – 9.
[8] 庄福振,罗平,何清,等. 迁移学习研究进展[J]. 软件学报,2015,26(1):26 – 39.
Zhuang Fuzhen,Luo Ping,He Qing,et al. Survey on transfer learning research[J]. Journal of Software,2015,26(1):26 – 39. (in Chinese)
[9] 王皓,高阳,陈兴国. 强化学习中的迁移:方法和进展[J]. 电子学报,2008,36(z1):39 – 43.
Wang Hao, Gao Yang, Chen Xingguo. Transfer of reinforcement learning:the state of the art[J]. Acta Electronica Sinica,2008,36(z1):39 – 43. (in Chinese)
[10] Sorg J,Singh S. Transfer via soft homomorphisms[A]. Proc of The 8th International Conference on Autonomous Agents and Multi-agent Systems Richland(AAMS 2009)[C]. Budapest Hungary,2009. 2:741 – 748.
[11] Lazaric A,Restelli M,Bonarini A. Transfer of samples in batch reinforcement learning[A]. Proc of the 25th International Conference on Machine Learning[C]. NewYork,ACM. 2008. 1:544 – 551.
[12] Castro P S,Precup D. Using bisimulation for policy transfer in MDPs[A]. Proc of the 9th International Conference on Autonomous Agents and Multiagent Systems(AAMS 2010)[C]. Toronto: IFAAMS,2010. 1:1399 – 1400.
[13] Ferns N,Castro P S,Precup D,et al. Methods for computing state similarity in Markov decision processes[J]. Computer Science,2012:174 – 181.
[14] Busoniu L,Babuska R,Schutter B D,et al. Reinforcement Learning and Dynamic Programming Using Function Approximators[M]. CRC Press,New York,USA,2010.
[15] Givan R,Dean T and Greig M. Equivalence notions and model minimization in markov decision processes[J]. Artificial Intelligence,2003,147(1 – 2): 163 – 223.
[16] Devolder P,Janssen J,Manca R,et al. Markov Chains[M]. Basic Stochastic Processes. John Wiley & Sons,Inc,2015. 77 – 112.

作者简介

朱 斐 男,1978 年出生,江苏苏州人,博士,苏州大学副教授,主要研究方向为机器学习,人工智能,生物信息学等.
E-mail: zhufei@suda.edu.cn

刘 全(通信作者) 男,1969 年出生,内蒙古牙克石人,博士后,苏州大学教授、博士生导师,主要研究方向为强化学习、人工智能、自动推理等.
E-mail: quanliu@suda.edu.cn

傅启明 男,1985 年出生,江苏淮安人,博士,苏州科技大学讲师,主要研究方向为强化学习、人工智能等.
E-mail: fqm_1@126.com

陈冬火 男,1974 年出生,江西九江人,博士,苏州大学讲师,主要研究方向为程序分析和验证、模型检验、自动推理和机器学习等.
E-mail: dhchen@suda.edu.cn

伏玉琛 男,1968 年出生,江苏徐州人,博士,苏州大学教授、硕士生导师,主要研究方向为强化学习、人工智能等.
E-mail: yuchenfu@suda.edu.cn

王 辉 男,1968 年出生,陕西西安人,硕士,苏州大学讲师,主要研究方向为强化学习、人工智能等.
E-mail: huiw@suda.edu.cn