软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(11):2667−2675 [doi: 10.3724/SP.J.1001.2013.04471]

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

一种基于自生成样本学习的奖赏塑形方法*

钱 煜, 俞 扬, 周志华

(计算机软件新技术国家重点实验室(南京大学),江苏 南京 210023)

通讯作者: 俞扬, E-mail: yuy@nju.edu.cn, eyounx@gmail.com

摘 要: 强化学习通过从以往的决策反馈中学习,使 Agent 做出正确的短期决策,以最大化其获得的累积奖赏值.以往研究发现,奖赏塑形方法通过提供简单、易学的奖赏替代函数(即奖赏塑性函数)来替换真实的环境奖赏,能够有效地提高强化学习性能.然而奖赏塑形函数通常是在领域知识或者最优策略示例的基础上建立的,均需要专家参与,代价高昂.研究是否可以在强化学习过程中自动地学习有效的奖赏塑形函数.通常,强化学习算法在学习过程中会采集大量样本.这些样本虽然有很多是失败的尝试,但对构造奖赏塑形函数可能提供有用信息.提出了针对奖赏塑形的新型最优策略不变条件,并在此基础上提出了 RFPotential 方法,从自生成样本中学习奖赏塑形.在多个强化学习算法和问题上进行了实验,其结果表明,该方法可以加速强化学习过程.

关键词: 奖赏塑形;自生成样本;策略不变;强化学习

中图法分类号: TP181 **文献标识码**: A

中文引用格式: 钱煜,俞扬,周志华.一种基于自生成样本学习的奖赏塑形方法.软件学报,2013,24(11):2667−2675. http://www.jos.org.cn/1000-9825/4471.htm

英文引用格式: Qian Y, Yu Y, Zhou ZH. Shaping reward learning approach from passive samples. Ruan Jian Xue Bao/Journal of Software, 2013,24(11):2667−2675 (in Chinese). http://www.jos.org.cn/1000-9825/4471.htm

Shaping Reward Learning Approach from Passive Samples

QIAN Yu, YU Yang, ZHOU Zhi-Hua

(National Key Laboratory for Novel Software Technology (Nanjing University), Nanjing 210023, China)

Corresponding author: YU Yang, E-mail: yuy@nju.edu.cn, eyounx@gmail.com

Abstract: Reinforcement learning (RL) deals with long-term reward maximization problems via learning correct short-term decisions from on previous experience. It has been revealed that reward shaping, which provides simpler and easier reward functions to replace the actual environmental reward, is an effective way to guide and accelerate reinforcement learning. However, building a shaping reward requires either domain knowledge or demonstrations from an optimal policy, both involve participation of human experts that is costly. This work investigates whether it is possible to automatically learn a better shaping reward along with an RL process. RL algorithms commonly sample a lot of trajectories throughout the learning process. Those passive samples, though containing many failed attempts, may provide useful information for building a shaping reward function. A policy-invariance condition for reward shaping is introduced as a more effective way to handle noisy examples, followed by the RFPotential approach to learn a shaping reward from massive examples efficiently. Empirical studies on various RL algorithms and domains show that RFPotential can accelerate the RL process.

Key words: shaping reward; passive sample; policy-invariance; reinforcement learning

在强化学习中[1],一个 Agent 通过学习从环境状态到行为的映射,使其从环境中获得的累积奖赏值最大.没有任何先验知识的强化学习方法通常会遇到收敛速度慢的问题,其主要原因是学习过程中需要搜索的动作空间一般都比较大.一种直接且有效的解决方法就是将先验知识引入到学习过程中[2−4].

* 基金项目: 江苏省自然科学基金(BK2012303); 百度开放课题(181315P00651)
收稿时间: 2013-04-06; 修改时间: 2013-07-17; 定稿时间: 2013-08-27

奖赏塑形[5,6]是一种加速强化学习过程的有效方法.它通过提供一个可能失真但却易于学习的替代奖赏函数来引入先验知识,以使 Agent 可以更快地学习到原始问题的近似最优策略.Ng 等人[7]指出,任何不改变最优策略的奖赏塑形函数都可以被分解成基于状态的势函数之差,从而为奖赏塑形函数的设计提供了指导思想.随后,更多的不改变最优策略的势函数形式被进一步发现,包括基于状态和动作的势函数[1]和动态环境下基于值的势函数[8].Wiewiora[1]和 Devlin 等人[8]研究了奖赏塑形函数和初始化 Q 函数表的关系,他们发现:当势函数是静态时,两者是等价的;而当势函数是动态时则不等价.

另外一类相关的研究是逆强化学习(inverse reinforcement learning,简称 IRL).在这种设定下,Agent 可以观察到来自最优策略的行为样本,并从这些样本中估计出奖赏函数,然后进行传统的强化学习.Ng 和 Russell[9]提出了从最优行为样本中学习奖赏函数需要满足的约束条件,推动了很多逆强化学习算法(如最大熵方法[10]等)和成功应用(如车辆导航[11]等)的出现.

上述方法通过直接设计奖赏塑形函数或者观察最优行为样本来结合专家知识.然而在很多情况下,获取专家知识的代价高昂,使得这些方法的可行性受到限制.同时我们注意到,强化学习算法在学习过程中通常会采样很多样本,这自然地产生了一个问题:是否可以从这些免费的自生成样本中提取出有用的信息来加速学习过程?这个问题起初看上去似乎不可行,因为强化学习算法本身就是从这些样本中学习.然而现有的强化学习算法通常只利用了采样中的部分信息,如果换一个角度考察这些样本,应当可以更充分地利用其中的信息.

由于自生成样本并不是来自于最优策略(否则就已经得到了一个最优策略),不满足逆强化学习算法中的约束条件,难以利用逆强化学习估计奖赏函数,因此,本文提出了基于奖赏塑形的思路,将自生成样本中成功的尝试和失败的尝试进行对比,通过监督学习的方法找到一些有潜力的状态,并利用奖赏塑形函数鼓励 Agent 更多地访问这些潜力状态以加速学习过程.实现该思路主要需要攻克两个难点:首先,自生成样本有很多噪音,为此,首先,本文给出一个针对奖赏塑形的最优策略不变条件,并在此基础上提供一种有效的方法来处理这些噪音样本;其次,自生成样本数量巨大,这就要求所提出的方法能够快速地学习和更新,以便在实际中可用.本文使用完全随机森林[12]作为监督学习算法,形成了本文提出的 RFPotential 方法.在多种设置下的实验结果初步验证了利用自生成样本学习的可能性.

本文第 1 节介绍相关工作.第 2 节介绍本文方法 RFPotential.第 3 节介绍实验方法和结果.第 4 节总结全文.

1 相关工作

1.1 强化学习

在标准的强化学习设置下,一个 Agent 的目标是通过与环境的相互作用,学习一个关于动作选择的最优策略以使其获得的累积奖赏最大.通常,强化学习问题通过马尔可夫决策过程(Markov decision process,简称 MDP)建模.马尔可夫决策过程由一个四元组(S,A,P,R)组成,其中,S 是状态集合,A 是动作集合,$P(s'|s,a)$是在状态 s 采取动作 a 转移到状态 s'的概率,$R(s,s')$则表示从状态 s 迁移到状态 s'所得到的瞬时奖赏.一个策略通常被定义成从状态集合到动作集合的映射函数,如$\pi:S\to A$,而 Agent 的目标是学习到一个最大化长期奖赏的策略π.这里的奖赏可以是期望折扣奖赏 $E_\pi\left[\sum_{t=0}^{\infty}\gamma^t r_{t+1}\right]$和 T 步平均奖赏 $E_\pi\left[\frac{1}{T}\sum_{t=0}^{T-1}r_{t+1}\right]$,其中,$\gamma\in[0,1]$,$r_{t+1}$表示由 R 决定的瞬时奖赏.

一类经典的强化学习算法是基于值的方法,它将注意力放在如何学习关于状态和动作的值函数上.在有折扣的情况下,一个策略π在状态 s 的值函数可以表示为 $V^\pi(s)=E_\pi\left[\sum_{t=0}^{\infty}\gamma^t r_{t+1}\,|\,s_0=s\right]$,而状态-动作值函数可以表示为 $Q^\pi(s,a)=E_\pi\left[\sum_{t=0}^{\infty}\gamma^t r_{t+1}\,|\,s_0=s,a_0=a\right]=E_{s'\sim P(\cdot|s,a)}[R(s,s')+\gamma V^\pi(s')]$.最优状态值函数是 $V^*(s)=\max_\pi V^\pi(s)$,最优动作-状态值函数是 $Q^*(s,a)=\max_\pi Q^\pi(s,a)$,通过贝尔曼方程可以得到:$V^*(s)=\max_a E_{s'\sim P(\cdot|s,a)}[R(s,s')+\gamma V^*(s')]$,$Q^*(s,a)=E_{s'\sim P(\cdot|s,a)}[R(s,s')+r\max_{a'}Q^*(s',a')]$.

时间差分方法是一类不基于模型的,用于解决上述贝尔曼方程的算法.Sarsa 和 Q-Learning 算法是两种典型

的时间差分算法.Sarsa 算法的一步更新可以表示成:

$$Q(s,a)=(1-\alpha)Q(s,a)+\alpha[R(s,s')+\gamma Q(s',a')].$$

Q-Learning 算法的一步更新可以表示成:

$$Q(s,a)=(1-\alpha)Q(s,a)+\alpha[R(s,s')+\gamma\max_{a'}Q(s',a')].$$

目前,很多时间差分算法的变化版本被用来处理连续状态空间的问题.最小二乘时间差分算法(least-squares temporal difference,简称 LSTD)[13−15]通过线性组合一组基函数来直接近似值函数.由于运行速度快和鲁棒性好,LSTD 算法得到了广泛的应用.

与值函数法不同的一类方法是策略梯度[16]方法,它通过在一个固定的策略集合中使用梯度下降方法和策略搜索技术来最大化期望累积奖赏.策略梯度方法基于一个参数化的策略π_θ,其中,参数θ决定状态 s 下的动作选择 a.通常情况下,对于离散问题,使用吉布斯策略[16]: $\pi_\theta(a\,|\,s)=\exp(\phi(s,a)^\top\theta)\Big/\sum_b\exp(\phi(s,b)^\top\theta)$; 而对于连续问题,使用高斯策略[17,18]: $\pi_\theta(a\,|\,s)=F(\phi(s,a)^\top\theta_1,\theta_2)$. 非参数化策略梯度下降(non-parametirc policy gradient,简称 NPPG)[19]方法进一步提出,可以通过函数梯度的思想搜索一个非参数化策略.

1.2 奖赏塑形

奖赏塑形方法通过引入额外的奖赏来加速学习过程.通常情况下,新的奖赏函数表示成如下的形式:

$$R_s(s,s')=R(s,s')+F(s,s'),$$

其中,R 表示原始问题的奖赏函数,F 表示奖赏塑形函数.奖赏塑形函数通常根据先验领域知识构建而成,并且被证明在很多问题上都是非常有效的[4,6].

可是有研究发现[4],如果奖赏塑形方法使用不当,会导致非常差的性能,这主要因为最优策略由于奖赏塑形的加入而发生改变.Ng 等人[7]随后证明:把奖赏塑形函数分解成基于状态的势函数之差的形式:

$$F(s,s')=\gamma\Phi(s')-\Phi(s)$$

是最优策略保持不变的一个充分必要条件(其中,Φ可以是任意函数).Wiewiora[20]进一步指出:对 Q 值表进行初始化的方法等价于引入奖赏塑形函数.目前,已经有很多关于保证最优策略不变的奖赏塑形函数的形式的研究:Wiewiora 等人[1]指出,考虑动作的势函数是有效的,比如 $F(s,a,s',a')=\gamma\Phi(s',a')-\Phi(s,a)$;Devlin 和 Kudenko[8]指出,基于时间变化的势函数也是有效的,比如 $F(s,t,s',t+1)=\gamma\Phi(s',t+1)-\Phi(s,t)$,但是这种函数形式已经不等价于初始化 Q 值表的方法.所有上述发现都让对奖赏塑形函数的研究变得可行,但是这些方法都需要人工地设计一个有用的奖赏塑形函数.

2 RFPotential 方法

强化学习算法通常采样得到一组样本$\{(S_1,R_1),\ldots,(S_n,R_n)\}$,每一个样本由一组交替出现的状态和动作组成: $(S_i=s_0^{(i)},a_0^{(i)},s_1^{(i)},a_1^{(i)},\ldots,s_{m_i}^{(i)})$, R_i 表示从环境中得到的累积奖赏.

我们的想法是将累积奖赏高(好)的样本和累积奖赏低(坏)的样本进行比较,通过将这种比较形式化成一个监督学习问题,可以很容易地从中获得一些有用的信息,从而利用它们来指导如何自动地构建奖赏塑形函数.对状态层次的监督学习问题,本文把每一个样本拆包重组成一个新的训练数据集 $D=\{(s,y)|\forall i:s\in S_i,y=I(R_i>\eta)\}$,其中,$\eta$是事先设定的阈值.这个训练数据包括了状态和它对应的标记(I 是指示函数,当它内部的表达式为真时输出 1,否则输出 0),因此,可以利用监督学习算法学习近似后验概率模型 $\tilde{P}(y\,|\,s)$.

当建立了这样一个近似后验概率模型之后,就可以利用它构建奖赏塑形函数.此后,奖赏塑形就被用来鼓励 Agent 更频繁地访问那些好的状态,这就形成了本文的 RFPotential 方法.对于强化学习算法来说,RFPotential 方法相当于替换原始环境的奖赏函数:首先观察到强化学习算法的采样序列,然后构建奖赏塑形函数,并将原始环境中的奖赏替换成新的奖赏传递给强化学习算法.后面两节中将依次介绍如何实现一个有效的监督学习算法以及如何设计奖赏塑形函数.

2.1 学习后验概率

在近似值函数的研究中,存在一个非常普遍的设定:存在一些特征映射,将每一个状态向量映射到一个特征向量上.这样,为了简化起见,本文把前面的状态向量都当成特征向量.为了能够使用传统监督学习方法来有效地分析各个状态的好坏程度,通常需要大量的训练样本 D,因此采用的学习方法应该非常高效.

本文采用完全随机森林[12]来分析各个状态的好坏程度.完全随机森林由一组完全随机决策树组成,每棵树的生成过程如算法 1 所述.为了生成一棵完全随机决策树,新的训练数据集 D 被直接传递给树的根节点.在每一个节点 N 的构建过程中,正样本占所有样本的比例被首先记录到 N_c 中(第 2 行),然后,数据集被一个随机生成的超平面划分成两个子集(第 6 行~第 9 行),每个子集分别被用于生成 N 的左子节点和右子节点(第 10 行、第 11 行).当传递到节点的样本数量太少或者达到树的最大深度时,递归过程终止(第 3 行).可以发现,节点的建立过程非常快,因为它仅仅扫描 1 遍传递给它的所有样本而不需要任何优化过程.

算法 1. 完全随机决策树构造算法 CRT.

Input:

$D=\{(s_1,y_1),\ldots,(s_k,y_k)\}$:训练数据;$d$:当前深度;$q$:特征向量的维度.

Output:

N:树节点.

1. 创造一个节点 N
2. $N.c=|\{s_i|\forall i=1,\ldots,k:I(y_i=1)\}|/k$
3. **if** $|D|\leqslant 1$ or $d\leqslant 0$ or $N.c=1$ **then**
4. **return** N
5. **end if**
6. 产生一个随机 q 维向量 $w\sim\mathcal{N}(0,1)$
7. 令 θ 等于区间 $(\min_i w^\top s_i,\max_i w^\top s_i)$ 内的任意值
8. 令 $\mathcal{L}=\{(s,y)|\ \forall(s,y)\in\mathcal{D}:w^\top s<\theta\}$
9. 令 $\mathcal{R}=\{(s,y)|\ \forall(s,y)\in\mathcal{D}:w^\top s\geqslant\theta\}$
10. 递归调用决策树构造算法,$N.L=CRT(\mathcal{L},d-1,q)$
11. 递归调用决策树构造算法,$N.R=CRT(\mathcal{R},d-1,q)$
12. **return** N

在构建好一棵完全随机决策树之后,可以用它预测每一个状态的好坏程度:给定一个状态 s,让它从树的根节点开始访问这棵树,最终找到其最后到达的叶子节点.这个节点上所记录的正样本的比例 N_c 被用来近似状态 s 的后验概率.通常,由一棵随机决策树得到的对后验概率的估计是不太稳定的,因此,RFPotential 方法训练多棵完全随机决策树,最后它们输出的是平均预测值.总的来说,使用 T 棵树的完全随机森林估计后验概率的公式为

$$\tilde{P}(y=1|s)=\frac{1}{T}\sum_{t=1}^{T}N_c^{(t)}(s),$$

其中,$N^{(t)}(s)$表示状态 s 对应的第 t 棵树的叶子节点.

2.2 从后验概率得到奖赏塑形

基于上述的后验概率模型,本文提出 3 种奖赏塑形函数.在本节中,为简单起见,假设γ=1.

由于后验概率模型 $\tilde{P}(y=1|s)$ 表明一个状态属于一次成功尝试样本的概率,我们首先的想法就是构建一个奖赏塑形函数来鼓励 Agent 多去访问高后验概率的状态.根据 Ng 等人[7]提出的基于状态势函数的奖赏塑形函数框架,本文将第 1 个奖赏塑形函数 F_a 定义为

$$F_a(s,s')=\tilde{P}(y=1|s')-\tilde{P}(y=1|s).$$

当 Agent 的移动导致状态的后验概率上升时,它会收到一个额外的奖赏.

可是,估计的后验概率可能不是很准确,而且自生成样本显然不是根据最优策略得到的,也就是说,它们存

在很多噪音.所以对于监督学习算法而言,正确地学习后验概率是比较有难度的,完全依赖于可能错误的后验概率得到的奖赏函数 F_a 会向 Agent 提供错误的建议.

我们其次的想法是:仅仅在 Agent 移动到高后验概率的状态时才奖励它,这种赋值看起来更有说服力.首先定义阈值函数Θ:

$$\Theta_\theta(x)=\begin{cases}x, & x>\theta\\ 0, & x\leqslant\theta\end{cases}.$$

这样,第 2 个奖赏塑形函数 F_t 被定义为

$$F_t(s,s')=\Theta_\theta(\tilde{P}(y=1|s'))-\Theta_\theta(\tilde{P}(y=1|s)).$$

这个定义也满足基于状态势函数的框架,因此也不会影响最优策略.

F_t 的定义也存在一些弊端.在两个高后验概率状态间的移动可能会收到有噪音的奖赏,同时存在一些状态,尽管它的后验概率很低,但是却远远高于周围的大部分状态,那么对于这些状态的访问也应该被鼓励.为了找到一个好的奖赏塑形函数,本文定义了一类基于势函数的奖赏塑形函数:

定义 1. 一个基于动作最优的条件奖赏塑形函数可以被表示为

$$F(s,a,s')=\begin{cases}\Phi(s')-\Phi(s), & a\in A^*(s)\\ 0, & \text{otherwise}\end{cases},$$

其中,Φ是一个对所有从状态 s 采取动作 $A^*(s)$到达状态 s'的情况,均满足$\Phi(s')\geqslant\Phi(s)$的势函数.

定理 1. 给定一个 MDP $M=(S,A,P,R)$和定义 1 中的基于动作最优的条件奖赏塑形函数,令Π^*为 M 的最优策略的集合.对于新的 MDP $M'=(S,A,P,R')$,其中 $R'=R+F$,令Π'^*为 M'的最优策略的集合,则$\Pi'^*\subseteq\Pi^*$.

证明:对于 MDP M,最优值函数 V_M^* 满足:

$$V_M^*(s)=\max_a E_{s'\sim P(\cdot|s,a)}[R(s,s')+V_M^*(s')].$$

可以得到:

$$V_M^*(s)-\Phi(s)=\max_a E_{s'\sim P(\cdot|s,a)}[R(s,s')+\Phi(s')-\Phi(s)+(V_M^*(s')-\Phi(s'))].$$

定义 $V'_{M'}(s)=V_M^*(s)-\Phi(s)$. 注意到,对所有的最优动作来说,$F(s,a,s')=\Phi(s')-\Phi(s)$,可以得到:

$$V'_{M'}(s)=\max_a E_{s'\sim P(\cdot|s,a)}[R'(s,s')+V'_{M'}(s')].$$

这是一个对于 MDP M'的贝尔曼最优方程,因此肯定有唯一的最优值函数,即 $V_{M'}^*=V'_{M'}$. 注意到,动作-状态值函数可以写成:

$$Q_{M'}^*(s,a)=E_{s'\sim P(\cdot|s,a)}[R'(s,s')+V_{M'}^*(s')].$$

因此,当 $a\in A_M^*(s)$ 时,有:

$$Q_{M'}^*(s,a)=E_{s'}[R(s,s')-\Phi(s)+V_M^*(s')]=Q_M^*(s,a)-E_{s'}[\Phi(s)]\geqslant Q_M^*(s,a)-E_{s'}[\Phi(s')].$$

其中的不等号是由定义 1 中$\Phi(s')\geqslant\Phi(s)$得到的.当 $a\notin A_M^*(s)$ 时,可以得到:

$$Q_{M'}^*(s,a)=E_{s'}[R(s,s')-\Phi(s')+V_M^*(s')]=Q_M^*(s,a)-E_{s'}[\Phi(s')].$$

因此,对于所有的状态 s,所有的动作$a\in A_M^*(s)$和$b\notin A_M^*(s)$,由 $A_M^*(s)$ 的定义有:$Q_M^*(s,a)>Q_M^*(s,b)$,因而 $Q_{M'}^*(s,a)>Q_{M'}^*(s,b)$. 故 $A_{M'}^*\subseteq A_M^*$, 命题得证. □

上述基于动作最优的条件奖赏塑形函数可以仅在Agent采取最优动作时提供额外奖赏,因此它引入的噪音相对来说就比较少.不足的是,它需要知道最优动作集合,而在学习到最优策略之前我们只能近似地估计.因此,本文假设最优动作可以使前、后状态的后验概率得到明显的提升.基于此假设,可以得到一个近似的基于动作最优的条件奖赏塑形函数 F_d:

$$F_d(s,s')=\begin{cases}F_a(s,s'), & F_a(s,s')\geqslant\theta\\ 0, & \text{otherwise}\end{cases}.$$

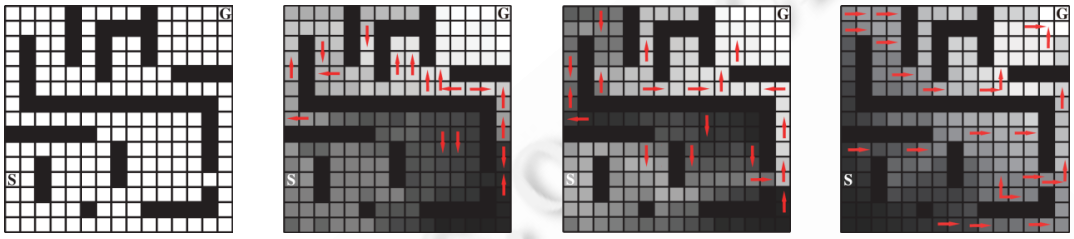
这是本文的第 3 个奖赏塑形函数.

3 实 验

本文方法 RFPotential 在 Maze 和 Mountain Car 这两个强化学习问题上与多种典型强化学习方法进行比较.在所有实验中,RFPotential 方法训练 20 棵最大深度为 10 的树组成的完全随机森林,每种算法都被重复运行 100 次,以取得平均性能.

3.1 Maze

本文首先采用一个简单的 15×15 的 Maze 问题,如图 1(a)所示,其中,S 表示初始位置,G 表示终止位置,所有白色的格子都是可以到达的,而黑色的格子则表示障碍物.在每一个格子中,Agent 从动作集合{向左,向右,向上,向下}选择一个动作,并以 0.9 的概率沿这个方向移动(除非目标位置是障碍物),同时,它也有 0.1 的概率任意选择一个方向移动.



(a) Maze 问题 (b) 收集 15 个样本 (c) 收集 40 个样本 (d) 收集 60 个样本

Fig.1 The 15×15 Maze problem and overlaps of the estimated posterior probability and the rewarding moves

图 1 15×15 的 Maze 问题和估计的后验概率及奖赏动作的显示

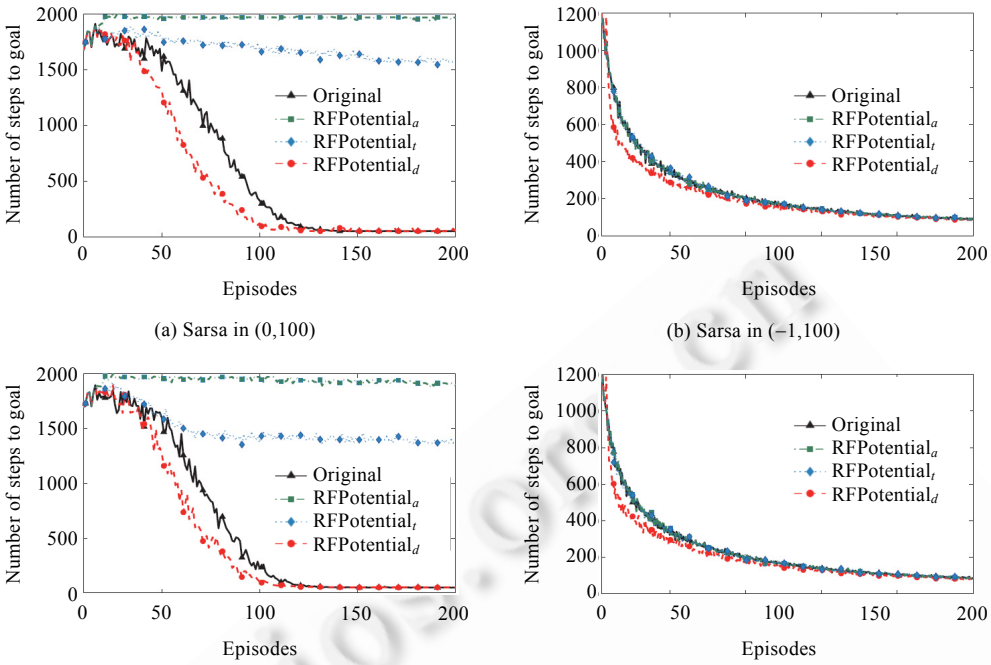
在本次实验中,本文采用 Sarsa 和 Q-Learning 算法作为原始强化学习算法,其中,折扣率γ=0.95,学习率α=0.1.对于两种算法,初始化时将 Q 值表中的所有位置都设置为 0 并且采用ε=0.1 的ε贪心算法.每次尝试均从初始位置 S 出发,在到达终止位置或者尝试步数超过 2 000 步时停止.本文采用两种不同的瞬时奖赏设置:(1) 对于到达非终止位置的移动,瞬时奖赏为 0;否则为 100;(2) 对于到达非终止位置的移动,瞬时奖赏为–1;否则为 100.在后文中,为简单起见,分别用(0,100)和(–1,100)表示.

为了能够使用 RFPotential 方法,状态需要用特征向量进行表示.本文将位置信息(x,y)进行二次平方展开得到的$(x,y,x^2,y^2,2xy)$作为其特征向量.RFPotential 在收集到 15 次尝试样本之后构造完全随机森林,并开始提供奖赏塑形.RFpotential 方法每收集到一次新的尝试样本,就更新完全随机森林,直到收集到 60 次尝试样本为止.在 RFPotential 方法 F_t 中,参数θ被设置成所有训练样本后验概率的平均值,而 F_d 中参数θ在奖赏设定(0,100)下被设置成 0.15,而在奖赏设定(–1,100)下被设置为 0.01.

实验的结果如图 2 所示,所有图的纵坐标表示的是从初始位置出发到达终止位置所需的平均步数.在相同的奖赏设定下可以看到,Sarsa 和 Q-Learning 的性能曲线基本一致.在实验设定(0,100)下可以看出,RFpotential 方法 F_a 和 F_t 导致了比原始算法更慢的收敛速率,而方法 F_d 则表现出更好的收率速度.在实验设定(–1,100)下,RFpotential 方法 F_a 和 F_t 对收敛速度几乎没有影响,只有 F_d 方法表现出更好的收敛速度.

图 1(b)~图 1(d)显示了在奖赏设定(0,100)下,采用 Q-Learning 作为原始强化学习算法,本文方法 RFpotential 估计的后验概率.通过将每个格子填充成不同的灰度,我们展示了每个状态对应的估计后验概率:越暗表示后验概率值越小,越亮则表示后验概率值越大.为了更好地可视化实验结果,我们将这些后验概率值进行直方图均衡化.箭头表示 F_d 方法中获得奖赏塑形的动作.在收集 15 次尝试样本以后,RFPotential 方法构建完全随机森林,第 1 次估计后验概率.从图 1(b)中可以看到,目标位置周围的区域都很亮,后验概率比较大,但一些箭头并不是所在状态的最优动作.在收集 40 次尝试样本后,大部分箭头都对应所在状态的最优动作;特别是在 Maze 右下角的需要穿越一扇“门”才能到达目标位置的情况中,箭头做出了正确的选择.在收集 60 次尝试样本后,估计的后验概率

从初始位置到目标位置形成了一个明显的灰度梯度,并且箭头全部都是最优动作(之一).



(a) Sarsa in (0,100) (b) Sarsa in (−1,100)

(c) Q-Learning in (0,100) (d) Q-Learning in (−1,100)

Fig.2 Experimental results on Maze

图 2 Maze 问题的实验结果

通过在 Maze 问题上的研究可以发现:首先,采用完全随机森林作为监督学习方法的 RFPotential 方法能够有效且准确地估计后验概率.正如前文所述:当样本数很少时,后验概率的估计很不准确;但当样本数量增加到一定程度时,估计的后验概率是比较可靠的;其次,尽管估计的后验概率是比较准确的,但它不一定总是有帮助的.塑像奖赏函数 F_d 是 3 个候选函数中唯一能够有效改善学习算法收敛效率的函数.

3.2 Mountain Car

Mountain Car 任务是一个典型的强化学习任务,其目标是驾驶一辆动力不足的小车到达一座陡峭的山的顶部.这个任务的状态由两个连续的变量表示:小车的位置[−1.2,0.5]和速度[−0.07,0.07],小车可以选择:向左加速、向右加速和不加速这 3 个不同的动作.没有达到目标位置的移动收到的瞬时奖赏为−1,而达到目标位置的瞬时奖赏为 100.

本文采用 Sarsa,Q-Learning,LSTD 和 NPPG 这 4 种强化学习算法作为原始强化学习算法.为了能够使用 Sarsa 和 Q-Learning 算法,在这两种情况下,本文将状态等间隔离散化,位置维度的区间大小为 0.1,而速度维度的区间大小为 0.01.在本次实验中,折扣率γ=1,学习率α=0.5.实验采用ε=0.1 的ε贪心算法,但ε的值随着尝试次数的增加而衰减,衰减因子是 0.99.对于采用 LSTD 作为原始算法的设定,最大步数被设置为 3 000,其他 3 种情况最大步数都为 1 000.方法 F_t 中,参数θ设置和 Maze 问题一样,而 F_d 中参数θ被设置成 0.03.

图 3(a)显示了采用 Sarsa 作为原始算法,各种方法达到目标位置所需的平均步数.可以看出:在尝试样本的数量较小时,F_t 和 F_d 方法都比原始方法收敛得快,但是当样本数增加时,F_t 方法变得很不稳定,效果也变得比原始方法糟糕;而 F_d 方法却和原始算法一样收敛到一个比较稳定的值.同时可以看到,F_a 方法收敛得很慢.

图 3(b)显示了采用 Q-Learning 作为原始算法,各个方法达到目标位置所需的平均步数.显然,F_a 和 F_t 方法都比 Q-Learning 方法收敛得更慢,只有 F_d 方法收敛速度稍微快一些.

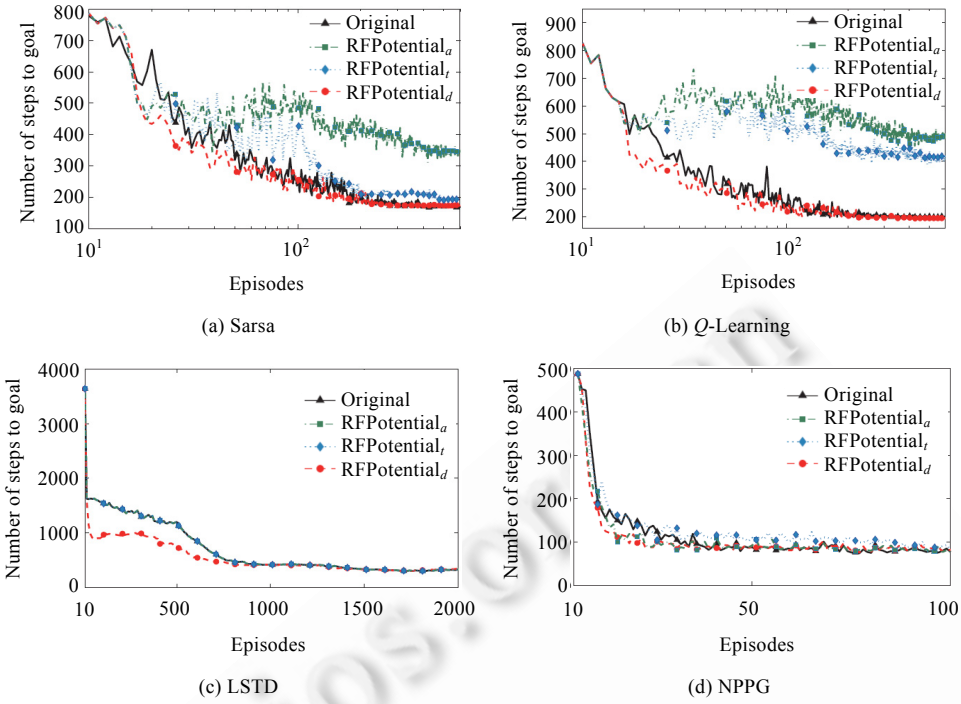


Fig.3 Experimental results on Mountain Car

图 3 Mountain Car 问题的实验结果

LSTD 算法直接处理状态空间连续的问题.图 3(c)表明:F_a 和 F_t 方法都没有对 LSTD 的收敛速度产生明显的影响,而 F_d 方法则在样本数量比较少的时候导致了非常明显的收敛速度的提高.

NPPG 算法是一种基于策略梯度的直接最大化累积奖赏的方法.图 3(d)表明:F_t 和 F_d 方法都比原始方法收敛得快;而 F_a 方法则导致原始算法收敛得稍微慢一些.

在上述所有实验中,尽管性能曲线各不相同,但是 F_d 方法都在加快收敛速度方面显示出其有效性.实验结果表明,从自生成样本中学习能够加速强化学习过程的奖赏塑形函数是可行的.

4 总 结

之前关于奖赏塑形函数的研究通常都需要领域知识或者来自最优策略的示例,本文研究能否从强化学习算法学习过程中产生的大量自生成样本中自动地学习有用的奖赏塑形函数.本文提出的 RFPotential 方法采用完全随机森林作为基本的监督学习方法,将样本中的成功尝试和失败尝试进行比较,最终建立后验概率模型.基于此模型,本文尝试了 3 种不同形式的奖赏塑形函数,其中两个遵循 Ng 等人[14]提出的基于状态势函数之差的框架,最后一个则满足本文提出的基于动作最优的条件奖赏塑形函数框架.在多种强化学习算法和问题上的实验结果表明,本文提出的基于动作最优的条件奖赏塑形函数能够加速强化学习过程,验证了从自生成样本中学习奖赏塑形函数的可行性.今后我们将进一步研究如何找到更好的基于动作最优的条件奖赏塑形函数和发现更多的最优策略不变性条件.

References:

[1] Wiewiora E, Cottrell GW, Elkan C. Principled methods for advising reinforcement learning agents. In: Proc. of the 20th Int'l Conf. on Machine Learning. Menlo Park: AAAI Press, 2003. 792−799.

[2] Babes M, Munoz de Cote E, Littman ML. Social reward shaping in the prisoner's dilemma. In: Proc. of the 7th Int'l Joint Conf. on Autonomous Agents and Multi-Agent Systems, Vol.3. Milton Keynes: IFAAMAS, 2008. 1389−1392.

[3] Marthi B. Automatic shaping and decomposition of reward functions. In: Proc. of the 24th Int'l Conf. on Machine Learning. New York: ACM Press, 2007. 601−608.

[4] Randlv J, Alstrm P. Learning to drive a bicycle using reinforcement learning and shaping. In: Proc. of the 15th Int'l Conf. on Machine Learning. New York: Morgan Kaufmann Publishers, 1998. 463−471.

[5] Dorigo M, Colombetti M. Robot shaping: Developing autonomous agents through learning. Artificial Intelligence, 1994,71(2): 321−370. [doi: 10.1016/0004-3702(94)90047-7]

[6] Mataric MJ. Reward functions for accelerated learning. In: Proc. of the 11th Int'l Conf. on Machine Learning. New York: Morgan Kaufmann Publishers, 1994. 181−189.

[7] Ng AY, Harada D, Russell SJ. Policy invariance under reward transformations: Theory and application to reward shaping. In: Proc. of the 16th Int'l Conf. on Machine Learning. New York: Morgan Kaufmann Publishers, 1999. 278−287.

[8] Devlin S, Kudenko D. Dynamic potential-based reward shaping. In: Proc. of the 11th Int'l Joint Conf. on Autonomous Agents and Multiagent Systems. Milton Keynes: IFAAMAS, 2012. 433−440.

[9] Ng AY, Russell SJ. Algorithms for inverse reinforcement learning. In: Proc. of the 17th Int'l Conf. on Machine Learning. New York: Morgan Kaufmann Publishers, 2000. 663−670.

[10] Ziebart BD, Maas AL, Bagnell JA, Dey AK. Maximum entropy inverse reinforcement learning. In: Proc. of the 23rd AAAI Conf. on Artificial Intelligence. Menlo Park: AAAI Press, 2008. 1433−1438.

[11] Abbeel P, Dolgov D, Ng AY, Thrun S. Apprenticeship learning for motion planning with application to parking lot navigation. In: Proc. of the IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. Los Alamitos: IEEE Computer Society Press, 2008. 1083−1090. [doi: 10.1109/IROS.2008.4651222]

[12] Liu FT, Ting KM, Yu Y, Zhou ZH. Spectrum of variable-random trees. Journal of Artificial Intelligence Research, 2008,32(1): 355−384. [doi: 10.1613/jair.2470]

[13] Boyan JA. Technical update: Least-Squares temporal difference learning. Machine Learning, 2002,49(2-3):233−246. [doi: 10.1023/A:1017936530646]

[14] Bradtke SJ, Bartol AG. Linear least-squares algorithms for temporal difference learning. Machine Learning, 1996,22(1-3):33−57. [doi: 10.1023/A:1018056104778]

[15] Xu X, He H, Hu D. Efficient reinforcement learning using recursive least-squares methods. Journal of Artificial Intelligence Research, 2002,16(10):259−292. [doi: 10.1613/jair.946]

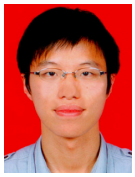
[16] Sutton RS, McAllester DA, Singh SP, Mansour Y. Policy gradient methods for reinforcement learning with function approximation. In: Advances in Neural Information Processing Systems 12. Cambridge: The MIT Press, 1999. 1057−1063.

[17] Peters J, Schaal S. Reinforcement learning of motor skills with policy gradients. Neural Networks, 2008,21(4):682−697. [doi: 10.1016/j.neunet.2008.02.003]

[18] Williams RJ. Simple statistical gradient following algorithms for connectionist reinforcement learning. Machine Learning, 1992, 8(3):229−256. [doi: 10.1007/BF00992696]

[19] Kersting K, Driessens K. Non-Parametric policy gradients: A unified treatment of propositional and relational domains. In: Proc. of the 25th Int'l Conf. on Machine Learning. New York: ACM Press, 2008. 456−463. [doi: 10.1145/1390156.1390214]

[20] Wiewiora E. Potential-Based shaping and *Q*-value initialization are equivalent. Journal of Artificial Intelligence Research, 2003, 19(1):205−208. [doi: 10.1613/jair.1190]



钱煜(1990－),男,江苏姜堰人,硕士生,主要研究领域为机器学习,强化学习.
E-mail: qiany@lamda.nju.edu.cn

周志华(1973－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为人工智能,机器学习,数据挖掘.
E-mail: zhouzh@nju.edu.cn

俞扬(1982－),男,博士,讲师,CCF 会员,主要研究领域为人工智能,演化计算,强化学习.
E-mail: yuy@nju.edu.cn