基于值函数迁移的启发式 Sarsa 算法

陈建平[1,2,3]，杨正霞[1,2,3]，刘全[4]，吴宏杰[1,2,3]，徐杨[5]，傅启明[1,2,3]

（1. 苏州科技大学电子与信息工程学院，江苏 苏州 215009；2. 苏州科技大学江苏省建筑智慧节能重点实验室，江苏 苏州 215009；
3. 苏州科技大学苏州市移动网络技术与应用重点实验室，江苏 苏州 215009；4. 苏州大学计算机科学与技术学院，江苏 苏州 215000；
5. 浙江纺织服装职业技术学院信息工程学院，浙江 宁波 315000）

摘 要：针对 Sarsa 算法存在的收敛速度较慢的问题，提出一种改进的基于值函数迁移的启发式 Sarsa 算法（VFT-HSA）。该算法将 Sarsa 算法与值函数迁移方法相结合，引入自模拟度量方法，在相同的状态空间和动作空间下，对新任务与历史任务之间的不同状态进行相似性度量，对满足条件的历史状态进行值函数迁移，提高算法的收敛速度。此外，该算法结合启发式探索方法，引入贝叶斯推理，结合变分推理衡量信息增益，并运用获取的信息增益构建内在奖赏函数作为探索因子，进而加快算法的收敛速度。将所提算法用于经典的 Grid World 问题，并与 Sarsa 算法、Q-Learning 算法以及收敛性能较好的 VFT-Sarsa 算法、IGP-Sarsa 算法进行比较，实验表明，所提算法具有较快的收敛速度和较好的稳定性。

关键词：强化学习；值函数迁移；自模拟度量；变分贝叶斯

中图分类号：TP391

文献标识码：A

doi: 10.11959/j.issn.1000−436x.2018133

Heuristic Sarsa algorithm based on value function transfer

CHEN Jianping[1,2,3], YANG Zhengxia[1,2,3], LIU Quan[4], WU Hongjie[1,2,3], XU Yang[5], FU Qiming[1,2,3]

1. Institute of Electronics and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China
2. Jiangsu Province Key Laboratory of Intelligent Building Energy Efficiency, Suzhou University of Science and Technology, Suzhou 215009, China
3. Suzhou Key Laboratory of Mobile Networking and Applied Technologies, Suzhou University of Science and Technology, Suzhou 215009, China
4. School of Computer Science and Technology, Soochow University, Suzhou 215000, China
5. Institute of Information Engineering, Zhejiang Fashion Institute of Technology College, Ningbo 315000, China

Abstract: With the problem of slow convergence for traditional Sarsa algorithm, an improved heuristic Sarsa algorithm based on value function transfer was proposed. The algorithm combined traditional Sarsa algorithm and value function transfer method, and the algorithm introduced bisimulation metric and used it to measure the similarity between new tasks and historical tasks in which those two tasks had the same state space and action space and speed up the algorithm convergence. In addition, combined with heuristic exploration method, the algorithm introduced Bayesian inference and used variational inference to measure information gain. Finally, using the obtained information gain to build intrinsic reward function model as exploring factors, to speed up the convergence of the algorithm. Applying the proposed algorithm to the traditional Grid World problem, and compared with the traditional Sarsa algorithm, the Q-Learning algorithm, and the VFT-Sarsa algorithm, the IGP-Sarsa algorithm with better convergence performance, the experiment results show that the proposed algorithm has faster convergence speed and better convergence stability.

Key words: reinforcement learning, value function transfer, bisimulation metric, variational Bayes

---

收稿日期：2018−03−22；**修回日期**：2018−07−13

通信作者：傅启明，fqm_1@126.com

基金项目：国家自然科学基金资助项目（No.61502329, No.61772357, No.61750110519, No.61772355, No.61702055, No.61672371, No.61602334）；江苏省自然科学基金资助项目（No.BK20140283）；江苏省重点研发计划基金资助项目（No.BE2017663）；江苏省高校自然科学基金资助项目（No.13KJB520020）；苏州市应用基础研究计划工业部分基金资助项目（No.SYG201422）

Foundation Items: The National Natural Science Foundation of China (No.61502329, No.61772357, No.61750110519, No.61772355, No.61702055, No.61672371, No.61602334), The Natural Science Foundation of Jiangsu Province (No.BK20140283), The Key Research and Development Program of Jiangsu Province (No.BE2017663), High School Natural Science Foundation of Jiangsu Province (No.13KJB520020), Suzhou Industrial Application of Basic Research Program Part (No.SYG201422)

1 引言

强化学习（RL, reinforcement learning）又称激励学习、增强学习，是在未知、动态环境中通过 agent 与环境的交互实现从状态到动作的映射，并获得最大期望累计奖赏的一类在线学习方法[1]。在强化学习问题中，新的强化学习任务与历史任务之间会存在某种相似性，因此可利用两者之间的相似性来提高目标任务的学习速率，这需要运用迁移学习（TL, transfer learning）方法。1995 年，迁移学习被首次以“learning to learn”的概念提出，引起学术界的广泛关注[2]。迁移学习主要包括 3 个方面：迁移什么、如何进行迁移、何时进行迁移。通过这 3 个方面，可以使迁移学习达到提高目标任务收敛速度的目的。然而迁移学习是对以往任务中学习的经验进行利用，从而提高目标任务的学习速率，但对于强化学习任务而言，其本身长期存在着平衡探索与利用之间关系的问题，有效地解决探索问题使 agent 获得最大化环境信息的轨迹，可以加快目标任务的学习速率。

近年来，迁移学习在强化学习领域已引起广大研究学者的关注。Ammar 等[3]通过优化不同任务间可转移的知识库，并通过对该知识库间不同任务构建映射关系，使新任务快速收敛。Gupta 等[4]通过构建状态空间到不变特征空间之间的映射关系，将知识映射到不变特征空间，并利用构建的映射关系实现知识的迁移，从而加快新任务的收敛速度。Laroche 等[5]在假设不同任务具有相同状态空间与动作空间的基础上，通过添加探索因子构建新的奖赏函数，实现不同任务间的知识迁移，提高算法在后续任务中的收敛性能。Barreto 等[6]提出在环境动态性不变的前提下，对不同任务之间的奖赏函数进行迁移，从而加快算法的收敛速度。

目前，研究学者已广泛关注强化学习问题中目标任务本身的探索与利用问题。传统的强化学习采用 ε-greedy 策略和 ε-soft 选择方法，在选择策略时添加一定的信息探索。贝叶斯与强化学习的结合成为解决探索与利用问题的另一重要方法。1998 年，Dearden 等[7]首次将强化学习与贝叶斯结合，利用贝叶斯概率模型对强化学习中值函数进行建模并确定置信度，平衡探索与利用的关系。Guez 等[8]提出利用蒙特卡罗树搜索，通过贝叶斯最优规划，平衡探索与利用的关系，加快算法的收敛速度。Little 等[9]在假设无外部奖赏反馈的情况下，通过获取信息增益，结合贝叶斯估计，求解最优探索策略，提高算法收敛性能。Mansour 等[10]通过贝叶斯后验分布构造策略模型，平衡探索与利用的关系，提高算法在后续任务中的收敛性能。Vien 等[11]运用贝叶斯强化学习模型，结合分级子目标方法，平衡探索与利用的关系，加快算法收敛速度。Wu 等[12]提出一种新的蒙特卡罗树搜索贝叶斯强化学习方法，解决在线贝叶斯强化学习问题，平衡探索与利用之间的关系，从而提高算法的收敛性能。傅启明等[13]利用高斯过程对带参数值函数进行建模，根据贝叶斯推理，求解值函数的后验分布，获取动作的信息增益并结合值函数的期望值，选择相应的动作，解决探索和利用的平衡问题，加快算法的收敛速度。

本文针对经典的 Sarsa 算法存在收敛速度慢的问题，提出一种基于值函数迁移的启发式 Sarsa 算法（VFT-HSA）。针对经典 Sarsa 算法中值函数初始值的设定直接影响算法收敛速度的问题，VFT-HSA 算法引入知识迁移，利用自模拟度量的方法，构造目标任务与历史任务之间的度量关系，通过设定阈值，迁移历史任务中的最优值函数，提高算法的收敛速度。针对大量算法问题中探索与利用不平衡的问题，VFT-HSA 引入启发式探索方法，利用贝叶斯推理，结合变分推理衡量信息增益，附加内在奖赏函数，从而提高算法的探索性能，加快算法的收敛速度。将 VFT-HSA 应用于 Grid World 问题，实验结果表明，VFT-HSA 较其他算法具有更快的收敛速度和较好的稳定性。

2 相关理论

2.1 马尔可夫决策过程

马尔可夫决策过程（MDP, Markov decision process）可以用来对强化学习问题进行建模，一个 MDP 通常可以表示为一个四元组 $M=<S,A,R,P>$。其中，S 表示所有状态的集合，$s_t \in S$ 是在 t 时间步下的状态；A 表示所有动作的集合，$a_t \in A$ 是在 t 时间步下的动作；R 表示奖赏函数，$R(s_t,a_t)$ 是在状态 S_t 下采取动作 a_t 所获得的立即奖赏；P 表示状态转移函数，$p(s,a,s')$ 是在状态 $s \in S$ 下采用动作 $a \in A$ 转移到下一状态 $s' \in S$ 的概率。

强化学习通过 agent 与环境交互求解最优策略，以获取最大期望回报。其中，策略 π 是在状态 s 下采取动作 a 概率的映射，可分为确定策略与随机策

略。确定策略 π 为 s 状态下选取特定动作 a，即 $\pi: S \to A$；随机策略 π 为动作空间 A 上的概率密度函数，即 $\pi: S \times A \to [0,1]$。

在评估 MDP 策略 π 时，引入值函数的概念，具体分为状态值函数 $V^{\pi}(s)$ 和状态动作值函数 $Q^{\pi}(s,a)$。$V^{\pi}(s)$ 是在 s 状态下采取策略 π 获得的回报的期望值，$Q^{\pi}(s,a)$ 是在状态动作对下采取策略 π 获得的回报的期望值。$V^{\pi}(s)$ 和 $Q^{\pi}(s,a)$ 可表示为相应 Bellman 公式的不动点解，即

$$V^{\pi}(s) = \sum_{a \in A} \pi(s,a) \sum_{s' \in S} p(s,a,s') \left[R(s,a) + \gamma V^{\pi}(s') \right] \quad (1)$$

$$Q^{\pi}(s,a) = \sum_{s' \in S} p(s,a,s') [R(s,a) + \gamma \sum_{a' \in A} \pi(s',a') Q^{\pi}(s',a')] \quad (2)$$

强化学习通过获得最优的状态值函数和最优的状态动作值函数，从而选取最优策略 π^*，其中，值函数 $V^*(s)$ 和 $Q^*(s,a)$ 可表示为

$$V^*(s) = \max_{a \in A} \sum_{s' \in S} p(s,a,s') [R(s,a) + \gamma V^*(s')] \quad (3)$$

$$Q^*(s,a) = \sum_{s' \in S} p(s,a,s') [R(s,a) + \gamma \max_{a' \in A} Q^*(s',a')] \quad (4)$$

式(3)和式(4)也被称为 Bellman 最优方程。

2.2 Sarsa 算法

在强化学习算法中，Sarsa 算法能够在未知奖赏函数与状态转移函数的情况下，采用状态动作值迭代找到最优策略，是一种在线学习算法。在 Sarsa 算法学习过程中，当状态动作对被无数次访问时，Sarsa 以概率 1 收敛到最优策略以及最优状态动作值函数，且策略将在有限的时间步内收敛至贪心策略。然而，Sarsa 算法是一种保守算法，为了减少损失，在学习过程中会选择相对安全的动作，这使 Sarsa 算法在选取动作时缺乏一定的探索，进而使 Sarsa 算法收敛速度相对较慢。Sarsa 算法具体流程如算法 1 所示[1]。

算法 1　Sarsa 算法

1) 初始化：对于 $\forall (s,a) \in S \times A$，$Q_0(s,a) = 0$
2) repeat (对于每一个情节 n)
3)　　初始化状态 s
4)　　在状态 s 下，根据行为策略选择动作 a
5)　　repeat (对于情节中的每一步)
6)　　采取动作 a，获得立即奖赏 r 和后续状态 s'
7)　　在状态 s' 下，根据行为策略选择动作 a'
8)　$Q(s,a) = Q(s,a) + \alpha [r + \gamma Q(s',a') - Q(s,a)]$
9)　　$s \leftarrow s'$，$a \leftarrow a'$
10)　　end repeat
11) end repeat
12) 输出：Q 值函数

2.3 自模拟度量

2003 年，Givan 等[14]首次将自模拟关系引入 MDP，并利用自模拟关系度量不同 MDP 中状态 S 之间的距离。其自模拟关系可简单表述为：若 2 个状态之间满足自模拟关系，那么 2 个状态之间的最优值函数或最优动作可相互共享。

定义 1　自模拟（bisumulation）关系。若 $E \subseteq S \times S$ 满足自模拟关系，那么对于任意 $s', s'' \in S$，$s'Es''$ 满足下列性质：1) $\forall a \in A$，满足 $R(s',a) = R(s'',a)$；2) $\forall a \in A$，$\forall F \in S \sim E$，满足 $\sum_{k \in F} p^a_{s'}(k) = \sum_{k \in F} p^a_{s''}(k)$，其中，$S$~$E$ 表示状态 S 集合满足 E 的等价状态集合。若 $s', s'' \in S$ 满足自模拟关系，可记为 $s' \sim s''$。

定义 2　度量（metric）。在状态 S 集合上的半度量 $d: S \times S \to [0,\infty)$，若 $\forall s', s'', s''' \in S$，则满足性质：1) $s' = s'' \Rightarrow d(s',s'') = 0$；2) $d(s',s'') = d(s'',s')$；3) $d(s',s''') \leqslant d(s',s'') + d(s'',s''')$。若性质 1)的逆命题同时成立，则 d 被称为状态 S 集合上的度量。

对于任意 2 个状态，它们之间的自模拟关系是“是”或“非”的关系，要么满足自模拟关系，要么不满足自模拟关系，但在实际应用中，该方法太过于严苛。如果 2 个状态的奖赏分布与状态转移概率分布极其近似，则 2 个状态极其近似，根据以上条件可推测 2 个状态具有相似的最优动作和最优值函数，但自模拟关系无法证明该推测。因而 Ferns 等[15]针对该问题，利用 Kantorovich 距离，提出衡量 2 个状态之间相似性关系的自模拟度量方法，并得到定理 1。

定理 1　D 为定义在状态集合 S 上的度量集合，且度量 $d \in D$。对于 $\forall s', s'' \in S$，定义 $J: D \to D$，$J(d)(s',s'') = \max_{a \in A} (d_a(s',s'') + \gamma T_n(d)(p(s',a), p(s'', a)))$，其中，$d_a(s',s'') = | p(s',a) - p(s'',a) |$，则 J 存在一个最小不动点 $d_{\sim}$，$d_{\sim}$ 是一个自模拟度量，$d_{\sim}(s',s'')$ 是状态 s' 和 s'' 之间的距离。

2.4 变分贝叶斯

变分贝叶斯最早由 Beal[16]提出，其可应用于隐

马尔可夫模型、混合因子分析、非线性动力学、图模型等。变分贝叶斯可较好地处理复杂统计模型。复杂统计模型由观测变量、未知参数和潜变量这 3 类变量组成，其中，未知参数和潜变量统称为不可观测变量。

采用变分贝叶斯具有如下优点：1)将不可观测变量的后验概率近似成其他变量，方便不可观测变量的推断；2)对于一个模型，给出边缘似然函数的下界，当边缘似然函数值最高时，表明模型拟合程度越好，通过该方法可获取最优模型。

定义 3 假设 X 为观测变量，Z 为 X 对应的潜变量，Θ 为未知参数集，根据贝叶斯推理，后验分布表示为

$$p(\Theta,Z\mid X)=\frac{p(\Theta)p(Z\mid X,\Theta)}{p(Z\mid X)} \tag{5}$$

变分贝叶斯采用更简单的近似分布 $q(\Theta)$ 来逼近真实的后验分布 $p(\Theta,Z\mid X)$，它们之间的联系通过最小化近似分布和后验分布之间的 Kullback-Leibler (KL)散度 $D_{\mathrm{KL}}\left[q(\Theta)\|p(\Theta,Z\mid X)\right]$ 来实现。

在变分贝叶斯中，给出边缘似然函数的下界 $L\left[q(\Theta)\right]$，通过最大化边缘似然函数值 $L\left[q(\Theta\mid Z)\right]$，获得最优模型解，可表示为

$$L\left[q(\Theta)\right]=\log p(Z\mid X,\Theta)-D_{\mathrm{KL}}\left[q(\Theta)\|p(\Theta,Z\mid X)\right] \tag{6}$$

3 VFT-HAS 算法思想及简介

3.1 值函数迁移

通常，对于 MDP，可以通过迭代方法求出最优状态值函数或最优动作值函数，再由最优值函数求解最优策略。但对于每一个 MDP，求解最优值函数都需要进行迭代计算，这会造成计算资源的浪费，因此考虑将已求解的历史最优值函数用于后续的 MDP 中，进而求解最优值函数。若 2 个状态相似，它们应该具有相似的最优状态值函数，并利用自模拟度量关系，对相似状态进行值函数迁移。在对值函数迁移方法进行介绍之前，先做如下假设。

假设 1 2 个不同 MDP（M_1 和 M_2）具有相同的状态集 S 和动作集 A，不同奖赏函数 R 和不同状态转移函数 P，即 $M_1=<S,A,R_1,P_1>$，$M_2=<S,A,R_2,P_2>$。

假设 1 中的 2 个不同 MDP，M_1 作为迁移值函数的原始 MDP，M_2 是被迁移的目标 MDP，s_1 和 s_2 分别表示 M_1 和 M_2 的状态，s_1' 和 s_2' 分别表示 M_1 和 M_2 的后续状态，V^*_1 和 V^*_2 分别表示 M_1 和 M_2 的最优状态值函数，π^*_1 和 π^*_2 分别表示 M_1 和 M_2 的最优策略。运用定理 2 说明 M_1 和 M_2 最优状态值函数之间与状态距离之间的关系。

定理 2 定义 D 为状态集合 S 上的度量集合，且度量 $d\in D$，$d_{\sim}(s_1,s_2)$ 是状态 s_1 和 s_2 之间的距离。若 $d_{\sim}(s_1,s_2)=\delta$，则 $|V_1^*(s_1)-V_2^*(s_2)|\leqslant\delta$。

关于定理 2 的证明可参考文献[17]，为了更加充分地说明定理 2，给出如下说明。

说明 对于 2 个已经确定的 MDP（M_1 与 M_2），2 个 MDP 之间遵循状态 S 和动作 A 共享的原则，其中，$s,s'\in S$，$a,b\in A$，则 agent 在学习过程中获得的立即奖赏与状态转移情况如图 1 所示。

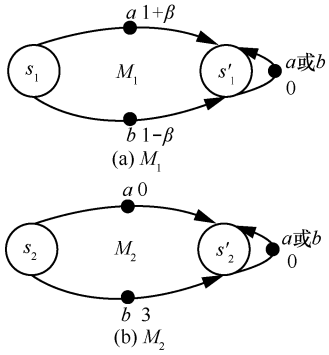


图 1 MDP 状态转移示意

在图 1 中，为了将不同的 MDP 中一样的状态区别开，采用 s_1 与 s_1' 分别表示 M_1 中的 s 和 s'，同样，采用 s_2 和 s_2' 分别表示 M_2 中的 s 和 s'。对 agent 获得的立即奖赏用动作旁的数值表示，如 1+β 表示在 s_1 状态下采取动作 a 到达下一状态 s_1' 获得的立即奖赏（$0\leqslant\beta\leqslant1$）。对于 M_1 而言，在状态 s_1 下的最优动作为 a，在 s_1 和 s_1' 下的最优值函数为 $V_1^*(s_1)=1+\beta$，$V_1^*(s_1')=0$。对于 M_2 而言，在状态 s_2 下的最优动作为 b，在 s_2 和 s_2' 下的最优值函数为 $V_2^*(s_2)=3$，$V_2^*(s_2')=0$。根据自模拟度量和图 1 的迁移条件，$d_{\sim}(s_1,s_2)=1+\beta$，$d_{\sim}(s_1',s_2)=3$，$d_{\sim}(s_1,s_2')=1+\beta$，$d_{\sim}(s_1',s_2')=0$（若 s_1' 和 s_2' 满足自模拟关系，依据定义 1，可认为 2 个状态是等价的）。可以获得如下关系式

$$|V_1^*(s_1)-V_2^*(s_2)|=1-\beta\leqslant d_{\sim}(s_1,s_2)=1+\beta$$

$$|V_1^*(s_1')-V_2^*(s_2')|=0\leqslant d_{\sim}(s_1',s_2')=0$$

$$|V_1^*(s_1')-V_2^*(s_2)|=3\leqslant d_{\sim}(s_1',s_2)=3$$

$$|V_1^*(s_1)-V_2^*(s_2')|=1+\beta\leqslant d_{\sim}(s_1,s_2')=1+\beta$$

由定理 2，给出不同 MDP 之间基于自模拟度量的值函数迁移算法，如算法 2 所示。

算法 2　基于自模拟度量的值函数迁移算法

输入　2 个 MDP（M_1 和 M_2），M_1 中的最优状态值函数 V_1^* 以及阈值参数 ς

输出　V_2

1) for $k=1$ to $k\leqslant|S_1|$ do
2)　for $m=1$ to $m\leqslant|S_2|$ do
3)　计算 $d_{\sim}(s_k,s_m)$
4)　end for
5) end for
6) for $i=1$ to $i\leqslant|S_2|$ do
7)　$s=\arg\min_{s\in S_1} d_{\sim}(s,s_i)$
8)　if $d_{\sim}(s,s_i)\leqslant\varsigma$ then
9)　$V_2(s_i)=V_1^*(s)$
10)　else
11)　$V_2(s_i)=0$
12)　end if
13) end for

3.2　基于变分贝叶斯的启发式探索

强化学习算法中经常使用启发式探索策略，如 ε-greedy 或 ε-soft，但该类启发式探索都依赖于非常低效的随机性的选择。而采用变分贝叶斯衡量信息增益，并构造内部奖赏函数，能有效地平衡强化学习中长期存在的探索与利用的问题。因此，本文提出一种新的基于变分贝叶斯的启发式探索方法。在介绍启发式探索之前，做如下定义。

定义 4　在 MDP 中，$p(s_{t+1}|\xi_t,a_t;\theta)$ 作为 agent 的环境动态模型随参数 $\theta\in\Theta$（Θ 为未知参数）变化，对于随参数 θ 变化的先验分布 $p(\theta)$，后验分布可通过贝叶斯方法表示为

$$p(\theta|\xi_t,a_t,s_{t+1})=\frac{p(\theta|\xi_t)p(s_{t+1}|\xi_t,a_t;\theta)}{p(s_{t+1}|\xi_t,a_t)} \tag{7}$$

其中，$\xi_t=\{s_1,a_1,s_2,\cdots,a_{t-1},s_t\}$ 为状态采样轨迹，s_{t+1} 为 t+1 时刻的状态。

在强化学习中，奖赏函数表示 agent 在每一个时间步获得的立即回报，而奖赏函数的好坏对 agent 的学习具有重要的作用，因此提出启发式探索的设计思路，将前后 2 个状态转移概率分布的信息增益作为内部奖赏。根据定义 4，agent 的环境动态模型 $p(s_{t+1}|\xi_t,a_t;\theta)$ 随参数 $\theta\in\Theta$ 变化，其后验分布 $p(\theta|\xi_t,a_t,s_{t+1})$ 随采样轨迹 $\xi_t=\{s_1,a_1,s_2,\cdots,a_{t-1},s_t\}$、动作 a_t、状态 s_{t+1} 更新。其中，在状态 s_t 下，动作 a_t 由行为策略选取，当行为策略保持不变时，后验分布 $p(\theta|\xi_t,a_t,s_{t+1})$ 与先验分布 $p(\theta|\xi_t)$ 近似。为了加大探索，使动作 a_{t+1} 分布不同于动作 a_t 分布，即后验分布 $p(\theta|\xi_t,a_t,s_{t+1})$ 不同于先验分布 $p(\theta|\xi_t)$，$p(\theta|\xi_t,a_t,s_{t+1})$ 与 $p(\theta|\xi_t)$ 的不同可用 KL 散度 $D_{\mathrm{KL}}\left[p(\theta|\xi_t,a_t,s_{t+1})\|p(\theta|\xi_t)\right]$ 表示。为了鼓励探索，应最大化 $D_{\mathrm{KL}}\left[p(\theta|\xi_t,a_t,s_{t+1})\|p(\theta|\xi_t)\right]$ 作为内部奖赏，因此 agent 奖赏函数可表示为 $r_{t+1}=r(s_t,a_t)+\eta D_{\mathrm{KL}}\left[p(\theta|\xi_t,a_t,s_{t+1})\|p(\theta|\xi_t)\right]$，其中，$\eta$ 为学习因子。然而后验分布难以获取，故结合变分贝叶斯推论，运用近似分布 $q(\Theta)$ 逼近后验分布。

定理 3　定义 $q(\Theta)$ 为近似概率分布集合，且 $q(\theta;\phi)\in q(\Theta)$，其中，$q(\theta;\phi)=\prod_{i=1}^{|\Theta|}N(\theta_i|\mu_i;\sigma_i^2)$，$\phi=\{\mu,\sigma\}$，$\sigma=\log(1+\mathrm{e}^{\rho})$，$\rho\in R$，$q(\theta;\phi)$ 是用于逼近后验分布 $p(\theta|\xi_t,a_t,s_{t+1})$ 的近似分布，为便于证明，将 ξ_t,a_t,s_{t+1} 用 D 表示，可表示为 $p(\theta|D)$。当 $q(\theta;\phi)=\mathrm{argmax}[\mathrm{E}_{\theta\sim p(\cdot;\theta)}\left[\log p(D|\theta)\right]-D_{\mathrm{KL}}\left[q(\theta;\phi)\|p(\theta|D)\right]]$ 时，则 $q(\theta;\phi)$ 近似于 $p(\theta|D)$。

证明　对于 $D_{\mathrm{KL}}\left[q(\theta;\phi)\|p(\theta|D)\right]$ 而言，当取最小值时，则 $q(\theta;\phi)$ 近似于 $p(\theta|D)$，即

$$\begin{aligned}
&D_{\mathrm{KL}}\left[q(\theta;\phi)\|p(\theta|D)\right]=\int q(\theta;\phi)\log\frac{q(\theta;\phi)}{p(\theta|D)}\mathrm{d}\theta\\
&=-\int q(\theta;\phi)\log\frac{p(\theta|D)}{q(\theta;\phi)}\mathrm{d}\theta\\
&=-\int q(\theta;\phi)\log\frac{p(\theta|D)}{q(\theta;\phi)p(D)}\mathrm{d}\theta\\
&=\int q(\theta;\phi)\left[\log q(\theta;\phi)+\log p(D)\right]\mathrm{d}\theta-\int q(\theta;\phi)\log p(\theta,D)\mathrm{d}\theta\\
&=\log p(D)+\int q(\theta;\phi)\log q(\theta;\phi)\mathrm{d}\theta-\int q(\theta;\phi)\log p(\theta,D)\mathrm{d}\theta
\end{aligned}$$

令 $L(q(\theta;\phi))=\int q(\theta;\phi)\log p(\theta,D)\mathrm{d}\theta-\int q(\theta;\phi)\log q(\theta;\phi)\mathrm{d}\theta$，$D_{\mathrm{KL}}\left[q(\theta;\phi)\|p(\theta|D)\right]=\log p(D)-L(q(\theta;\phi))$，由于对数 $\log p(D)$ 不依赖于不可观测变量，因此可用常量表示。Kullback-Leibler 散度关系如图 2 所示。

根据图 2 关系可知，为了得到 $q(\theta;\phi)$，使 KL 散度最小，可最大化 $L(q)$ 或最小化 $D_{\mathrm{KL}}(q\|p)-\log p(D)$。

证毕。

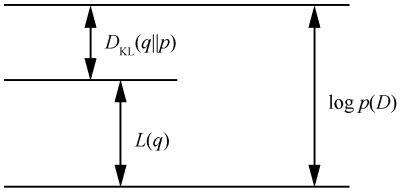


图 2 Kullback-Leibler 散度关系

定理 3 利用近似分布 $q(\theta;\phi)$ 逼近后验分布。根据以上定理构造奖赏函数如式(8)所示，进一步探索结构细节如图 3 所示。

$$r_{t+1}=r(s_t,a_t)+\eta D_{\mathrm{KL}}\left[q(\theta;\phi_{t+1})\|q(\theta;\phi_t)\right] \tag{8}$$

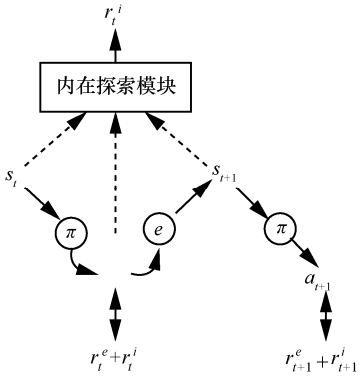


图 3 启发式探索结构

图 3 为启发式探索结构，其中，agent 执行当前策略 π 与环境 e 之间的交互作用，r_t^e 是环境 e 提供的外部奖赏 $r(s_t,a_t)$，r_t^i 是内部探索模块所提供的内在奖赏 $\eta D_{\mathrm{KL}}\left[q(\theta;\phi_{t+1})\|q(\theta;\phi_t)\right]$。

结合上述原理，给出一种改进的启发式内部奖赏函数 H 的更新式，如式(9)所示。

$$H(s_t,a_t,s_{t+1})=\begin{cases}\eta D_{\mathrm{KL}}\left[q(\theta;\phi_{t+1})\|q(\theta;\phi_t),\right], a_t=\pi^l(s_{t+1})\\ 0, \text{其他}\end{cases} \tag{9}$$

其中，$a_t=\pi^l(s_{t+1})$ 表示在状态 s_t 下选取的最优动作。其中，有

$$\phi_{t+1}=\arg\min\cdot \left[\underbrace{\overbrace{D_{\mathrm{KL}}\left[q(\theta\,|\,\phi_t)\|q(\theta\,|\,\phi_{t-1})\right]}^{\ell_{\mathrm{KL}}(q(\theta;\phi))}-\mathrm{E}_{\theta\sim p(:;\theta)}\left[\log p(s_t\,|\,\xi_t,a_t;\theta)\right]}_{\ell(q(\theta;\phi),s_t)}\right] \tag{10}$$

在式(10)中，梯度根据 KL 散度的曲率递减，为了优化式(10)，采用牛顿法计算 $D_{\mathrm{KL}}\left[q(\theta;\phi+\Delta\phi)\|q(\theta;\phi)\right]$（其中，$\Delta\phi=-\boldsymbol{H}^{-1}(\ell)\nabla_\phi\ell(q(\theta;\phi),s_t)$，$\boldsymbol{H}(\ell)$ 表示 $\ell(q(\theta;\phi),s_t)$ 的海森矩阵 $\boldsymbol{H}$(Hessian)）。假设 $q(\theta;\phi)$ 是可以进行因式分解的高斯分布，因此后验分布到先验分布的 KL 散度可表示为

$$D_{\mathrm{KL}}\left[q(\theta;\phi_t)\|q(\theta;\phi_{t+1})\right]=\frac{1}{2}\sum_{i=1}^{|\Theta|}\left(\left(\frac{\sigma_i^t}{\sigma_i^{t+1}}\right)^2+2\log\sigma_i^{t+1}-2\log\sigma_i^t+\frac{(\mu_i^{t+1}-\mu_i^t)^2}{(\sigma_i^{t+1})^2}\right)-\frac{|\Theta|}{2} \tag{11}$$

因为 KL 散度对参数用二次方表示，并且对数似然函数 $\log p(s_t\,|\,\xi_t,a_t;\theta)$ 相对于高度弯曲的 KL 项可看成是局部线性的，故 $\boldsymbol{H}$ 可直接近似计算 $\ell_{\mathrm{KL}}(q(\theta;\phi))$ 项的二阶导数。由于采用因式分解的高斯分布可近似表示为对角矩阵，故能有效提高计算速度。根据式(11)计算海森矩阵，其中，μ 和 ρ 项在 $\boldsymbol{H}$ 中的定义为

$$\frac{\partial^2\ell_{\mathrm{KL}}}{\partial\mu_i^2}=\frac{1}{\log^2(1+\mathrm{e}^{\rho_i})} \tag{12}$$

$$\frac{\partial^2\ell_{\mathrm{KL}}}{\partial\rho_i^2}=\frac{2\mathrm{e}^{2\rho_i}}{(1+\mathrm{e}^{\rho_i})^2}\frac{1}{\log^2(1+\mathrm{e}^{\rho_i})} \tag{13}$$

而其他项都为 0。此外，可通过二阶泰勒展开式 $\frac{1}{2}\Delta\phi\boldsymbol{H}\Delta\phi=\frac{1}{2}(\boldsymbol{H}^{-1}\nabla)^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{H}^{-1}\nabla)$ 近似 KL 散度，当 $\frac{1}{2}(\boldsymbol{H}^{-1}\nabla)^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{H}^{-1}\nabla)=0$ 时，可得散度最小值为

$$D_{\mathrm{KL}}\left[q(\theta;\phi+\Delta\phi)\|q(\theta;\phi)\right]\approx\frac{1}{2}\nabla_\phi\ell^{\mathrm{T}}\boldsymbol{H}^{-1}(\ell_{\mathrm{KL}})\nabla_\phi\ell \tag{14}$$

其中，$\boldsymbol{H}^{-1}(\ell_{\mathrm{KL}})$ 为对角线矩阵。

为了体现不同样本中 KL 散度的相对差异，对内部奖赏不直接采用 $D_{\mathrm{KL}}\left[q(\theta;\phi^{t+1})\|q(\theta;\phi^t)\right]$ 的形式，而将所有之前情节中 KL 散度的均值作为内部奖赏进行启发式探索，如式(15)所示。

$$D_{\mathrm{KL}}\left[q(\theta;\phi_{n+1}^{t+1})\|q(\theta;\phi_{n+1}^t)\right]=\frac{1}{N}\sum_{n=1}^{N}D_{\mathrm{KL}}\left[q(\theta;\phi_n^{t+1})\|q(\theta;\phi_n^t)\right] \tag{15}$$

3.3 VFT-HAS 简介

基于值函数迁移的启发式 Sarsa 算法主要利用自模拟度量方法对相似状态之间的以往值函数知识进行迁移，从而提高初始化值函数的精确性，并利用变分贝叶斯理论，获得信息增益作为内在奖赏函数进行启发式探索，结合 Sarsa 算法框架，利用 V-Q 算法中的更新方法更新值函数[18]，提高算法收敛速度，具体如算法 3 所示。

算法 3　基于值函数迁移的启发式 Sarsa 算法

输入　参数 δ、η，学习因子 α、β

输出　对于 $\forall s \in S$，策略 $\pi(s)=\arg\max_{a\in A} Q(s,a)$

1)初始化：利用算法 2 初始化状态值函数 V_0，且对于 $\forall(s,a)\in S\times A$，有 $Q_0(s,a)=0$

2) repeat (对于每一个情节 n)

3)　初始状态动作对 (s,a)

4)　repeat(对于情节中的每一个时间步 t)

5)　对 (s_t,a_t,s_{t+1}) 进行采样，获取相应转移概率和奖赏分别为 $p(s_{t+1}\mid s_t,a_t)$、$r(s_t,a_t)$

6)　$D_{\mathrm{KL}}\left[q(\theta;\phi_{n+1}^{t+1})\|q(\theta;\phi_{n+1}^{t})\right]=\frac{1}{2}\nabla_\phi \ell^{\mathrm{T}}\boldsymbol{H}^{-1}(\ell_{\mathrm{KL}})\nabla_\phi \ell$

7)　$D_{\mathrm{KL}}\left[q(\theta;\phi_{n+1}^{t+1})\|q(\theta;\phi_{n+1}^{t})\right]=\frac{1}{N}\sum_{n=1}^{N} D_{\mathrm{KL}}\left[q(\theta;\phi_{n}^{t+1})\|q(\theta;\phi_{n}^{t})\right]$

8)　$r'(s_t,a_t,s_{t+1})\leftarrow r(s_t,a_t)+\eta D_{\mathrm{KL}}\left[q(\theta;\phi_{n+1}^{t+1})\|q(\theta;\phi_{n+1}^{t})\right]$

9)　$V_n(s_t)=V_{n-1}(s_t)+\alpha(r'+\gamma V_{n-1}(s_{t+1})-V_{n-1}(s_t))$

10)　$Q_n(s_t,a_t)=Q_{n-1}(s_t,a_t)+\beta(r'+\gamma V_{n-1}(s_{t+1})-Q_{n-1}(s_t,a_t))$

11)　令 $s\leftarrow s_{t+1}$，且 $a\leftarrow a_{t+1}$

12)　end repeat

13)　if $\|Q_n-Q_{n-1}\|_\infty \leqslant \delta$ then

14)　算法终止

15)　end if

16)　$n=n+1$

17)　$q(\theta\mid\phi_{n+1})\leftarrow \arg\min(D_{\mathrm{KL}}\left[q(\theta\mid\phi_n)\|p(\theta)\right]-\mathrm{E}_{\theta\sim p(\cdot;\theta)}\left[\log p(s_{t+1}\mid\xi_t,a_t;\theta)\right])$

18) end repeat

基于值函数迁移的启发式 Sarsa 算法主要分为 3 个部分，第一部分利用算法 2 知识迁移进行初始化状态值函数；第二部分对状态和动作及下一个状态进行采样，通过变分贝叶斯理论衡量信息增益作为内部奖赏函数；第三部分在第二部分的基础上更新状态值函数和状态动作值函数，求解问题最优策略，提高算法学习速率。

4　实验及结果分析

为了研究算法的性能，将 VFT-HSA 应用在 Grid World 问题中，并针对算法收敛的速度以及算法的稳定性等方面进行分析，将 VFT-HSA 与 Sarsa 算法、Q-Learning 算法、VFT-Sarsa 算法[17]、IGP-Sarsa[19]算法在相同的实验环境中重复实验 24 次，取每次实验的平均值比较各算法的性能。

4.1　Grid World 问题介绍

采用的 Grid World 问题是 5×6 和 10×10 的格子世界问题，如图 4 所示。5×6 格子包含一个起始状态(0,4)、一个终止状态(5,0)和有障碍物的格子状态（灰色格子代表有障碍物）；10×10 格子包含一个起始状态(0,9)、一个终止状态(9,0)和有障碍物的格子状态（灰色格子代表有障碍物）。每个情节中，agent 从起始状态（S）出发，到达终止状态（T）。在任意状态下，agent 有 4 个动作可以选择，即 $\{a_0,a_1,a_2,a_3\}$，分别为上、下、左、右动作，如图 4 中箭头方向所示。agent 根据所选择的动作进行状态迁移，如果选择向上的动作，状态迁移到当前状态上面的状态；如果选择向下的动作，状态迁移到当前状态下面的状态。以 10×10 为例，当 agent 处于(0,9)状态，其选择向下的动作，则状态迁移至(0,8)；

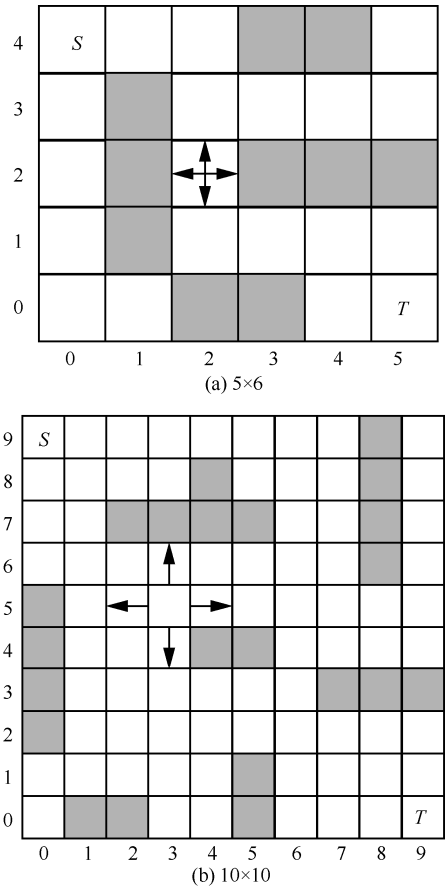
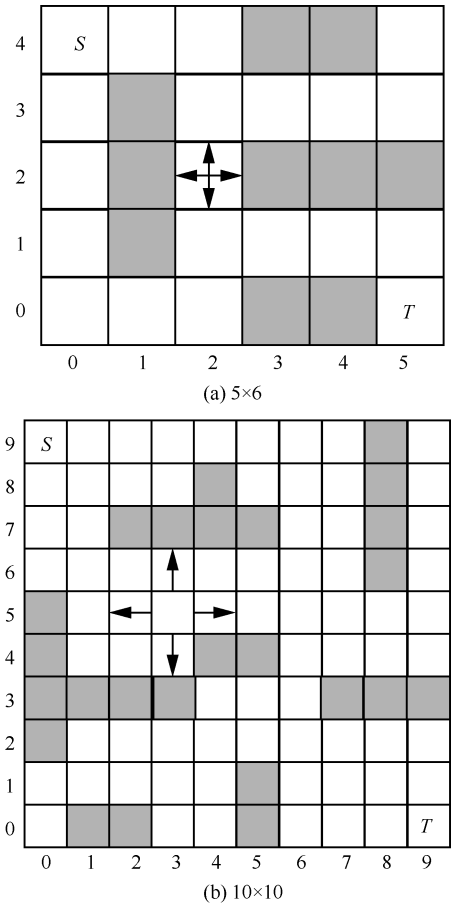


图 4　格子世界（目标 MDP）

当碰到障碍物或边界时，agent 留在原地不动。为了使问题更加复杂，假设 agent 将会以 0.1 的概率偶然地滑到与执行动作垂直的方向上。在状态迁移的过程中，到达终止状态时，agent 就可以获得一个较大的立即奖赏；碰到障碍物时，agent 获得一个最小的立即奖赏；其他情况下，agent 获得一个较小的立即奖赏。

4.2 实验设置

原始 MDP 如图 5 所示。若 2 个 MDP 具有相同的奖赏函数，则到达目标的立即奖赏为 1，碰到障碍物的立即奖赏为−1，其他状态的立即奖赏为−0.01，本次实验中情节数设为 400，每个情节中最大时间步数为 20 000，即当 agent 到达终止状态，则情节结束；或时间步达到 20 000 步，则情节结束。在相关参数中，折扣因子 γ=0.9，学习因子 $\alpha=\beta=\{0.03, 0.05, 0.3, 0.5\}$，参数 δ=0.001，η=0.55。



(a) 5×6

(b) 10×10

图 5 格子世界（原始 MDP）

4.3 实验分析

图 6 为算法 VFT-HSA、Sarsa、Q-Learning、VFT-Sarsa、IGP-Sarsa 在 5×6 的 Grid World 问题中的性能比较，其中，横坐标为情节数，纵坐标为 agent 收敛所需的平均步数。在实验过程中，每种算法都被独立执行 24 次，求出平均值，其中，值函数迁移其原始 MDP 如图 5 所示，目标 MDP 如图 4 所示。图 6 的 4 幅图分别为 4 个不同 α 下各算法的性能比较。学习因子 α 主要影响学习的收敛速度和收敛精度，一般情况下，α 越大，速度越快，相对精度越小，但 α 不宜过大，α 过大会使后期状态值产生振荡，影响收敛效果。对于 Sarsa、Q-Learning、VFT-Sarsa、IGP-Sarsa 算法，在图 6 中，当 α=0.03 和 α=0.05 时，算法收敛速度慢，收敛不稳定，200 个情节难以使 Sarsa、Q-Learning 算法收敛；当 α=0.3 和 α=0.5 时，算法的收敛速度明显加快，在 200 个情节内 Sarsa、Q-Learning 算法达到收敛，但算法收敛速度仍相对较低。对于 VFT-Sarsa、IGP-Sarsa 算法选取不同学习因子 α 时，收敛速度都有所提高，但对于本文提出的 VFT-HSA，在不同的 α 下，算法在 20 个情节左右均收敛至最优，具有更快的收敛速度和更好的收敛精度。这是因为算法学习过程中不但采用值函数迁移方法，使新任务下的初始值函数更接近最优值函数，而且采用启发式探索的方法，加大探索力度，提高算法的学习速率和稳定性。综上所述，VFT-HSA 选取不同的学习因子的值时，在保证收敛精度的情况下，进一步提高了算法的收敛速度，具有较强的顽健性。

图 7 为算法 VFT-HSA、Sarsa、Q-Learning、VFT-Sarsa、IGP-Sarsa 在 10×10 的 Grid World 问题中的性能比较，其中，横坐标为情节数，纵坐标为 agent 收敛所需的平均步数。为了验证算法在状态动作空间维度增大的情况下依旧具有较好的性能，将 5×6 的 Grid World 换成 10×10 的 Grid World，在实验过程中，每个算法都被独立执行 24 次，求出平均值，其中，值函数迁移其原始 MDP 如图 5 所示，目标 MDP 如图 4 所示。从图 7 可知，当 α=0.05 和 α=0.5 时，将 VFT-HSA 与 Sarsa、Q-Learning、VFT-Sarsa、IGP-Sarsa 算法进行比较，VFT-HSA 在保证收敛的情况下，不但有更快的收敛速度，而且具有更高的稳定性。Sarsa 算法在学习过程中随着状态动作空间维度的增大，算法不能保证较好的收敛；对于 Q-Learning 算法，当学习因子 α=0.05 时，并不能保证达到最终收敛状态，当 α=0.5 时，算法在 90 个情节左右时趋向收敛，但算法前期稳定性较差；对于 VFT-Sarsa、IGP-Sarsa 算法，当取不同的学习因子 α 时，算法在 50 个情节左右时趋于收

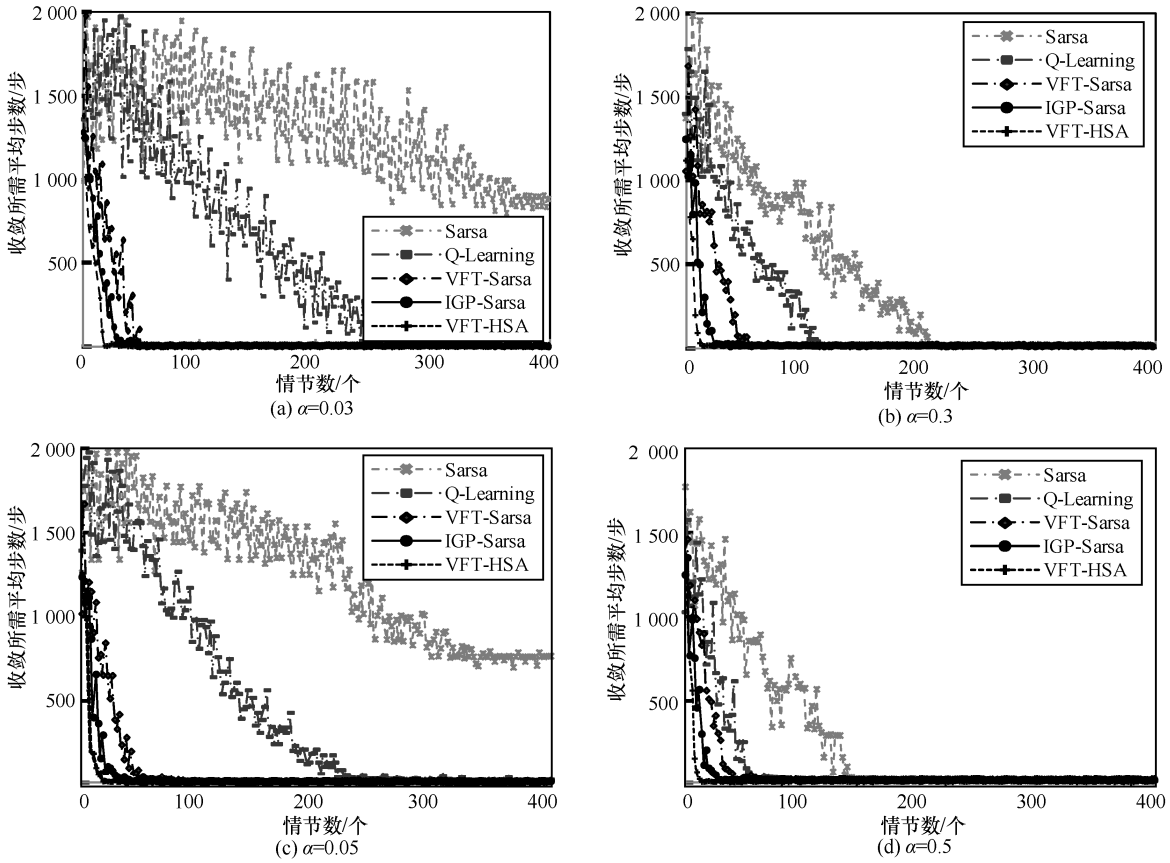


图 6　5×6 的 Grid World 问题中 5 种算法性能比较

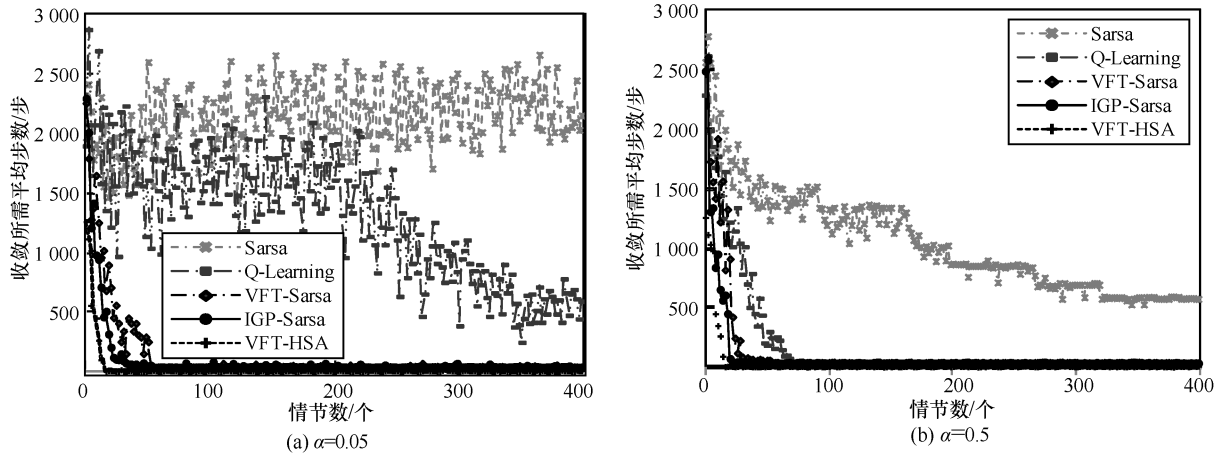


图 7　10×10 的 Grid World 问题中 5 种算法性能比较

敛，后期稳定性也较好，但与 VFT-HSA 相比，算法收敛速率仍相对较慢，算法稳定性也相对较差。因此，对于状态动作空间维度的增大，VFT-Sarsa、IGP-Sarsa 算法性能相对较差，VFT-HSA 性能相对较好，具有较好的顽健性。综上所述，VFT-HSA 在不同学习因子与不同状态动作空间维度问题中，都具有较快的收敛速度与较好的顽健性。

为了验证算法采用值函数迁移方法和启发式探索方法的收敛性能，图 8 分别表示 Sarsa 算法、本文提出的 VFT-HSA、不采用值函数迁移算法、不采用启发式探索算法在 10×10 的 Grid World 问题中达到收敛时所需的平均时间的变化趋势，其中，横坐标为情节数，纵坐标为情节结束后到达目标状态所需的时间。在实验过程中，每一个算法都独立执

行 24 次，取其平均值。在图 8 中，Sarsa 算法不能保证较好收敛，收敛性能较差；不采用值函数迁移算法在大约 40 个情节处收敛，而 VFT-HSA 在大约 30 个情节处收敛，VFT-HSA 相比于不采用值函数迁移算法收敛速度提升近 25%，因而不采用值函数迁移算法收敛速度较慢，这是因为不采用值函数迁移算法使算法运行过程中值函数的初始值未获得最优设置，算法收敛需要更多的样本数量，最终导致算法收敛速度慢；不采用启发式探索算法在大约 50 个情节处收敛，相比较而言，VFT-HSA 收敛速度提升近 40%，不采用启发式探索算法收敛性能不及 VFT-HSA，这是因为启发式探索算法在算法收敛过程中可以提供更多的启发式信息，加大 agent 探索力度，提高算法收敛速度。综上所述，在值函数迁移方法与变分贝叶斯启发式探索方法共同作用下，VFT-HSA 的收敛速度更快，收敛性能更好。

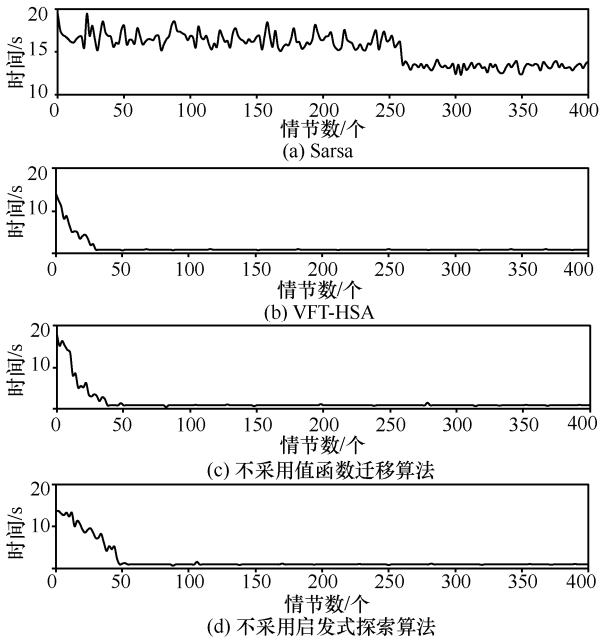


图 8 10×10 的 Grid World 问题中 4 种算法的性能比较

图 9 所示为 VFT-HSA 选择不同 η 值时，算法学习速率的变化，其中，横坐标为 η 的不同取值，纵坐标为算法收敛所需要的步数，图中所取的点是算法执行 24 次的平均值。由图 9 可知，不同的 Grid World 问题中，VFT-HSA 算法选取 η 值过小或过大，算法收敛所需平均步长都较大，算法的收敛性能较差。在 5×6 的 Grid World 问题中，当 η =0.5 左右时，平均步长在 36 左右达到收敛，算法收敛所需平均步长最少；在 10×10 的 Grid World 问题中，当 η =0.6 左右时，平均步长在 84 左右达到收敛，算法收敛所需平均步长最少。详细情况如表 1 所示。综上可知，η 取值不同造成算法的收敛性能有所不同，这是因为在不同的 Grid World 问题中，当 η 值设置得较小时，算法学习过程中对于额外获取的内部奖赏可以忽略不计，启发式探索方法无法起到加大探索力度的作用，从而使算法的收敛速度降低；当 η 值设置得较大时，会使 VFT-HSA 优先考虑额外获取的内部奖赏，对于外部奖赏忽略不计，这就不利于算法的收敛，所以选择合适的 η 值极其重要。合适的 η 值可以使 VFT-HSA 在启发式探索方法中达到最好的平衡点，从而大大提高算法的收敛速度与稳定性。同时，针对不同规模的 Grid World 问题，当算法收敛到最优时，η 的值基本一致，所以 VFT-HSA 算法对于 η 值具有较好的顽健性。

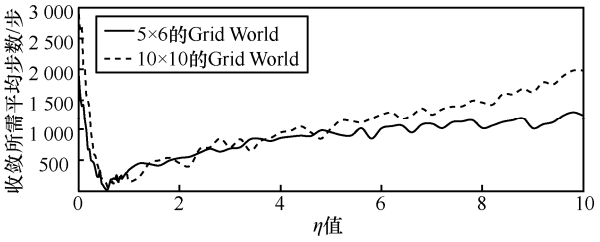


图 9 不同规模的 Grid World 问题中 VFT-HSA 取不同 η 值时收敛性能比较

表 1 不同规模的 Grid World 问题中 VFT-HSA 取不同 η 值时收敛所需平均步数比较

问题规模	η						
	0.3	0.5	0.6	0.8	2	5	8
5×6	456	36	102	236	543	956	1 026
10×10	612	156	84	156	456	1 023	1 456

5 结束语

本文针对 Sarsa 算法在维度较大的状态空间和动作空间的 MDP 中存在收敛速度慢的问题，提出一种改进的 VFT-HSA。在不同任务间具有相同状态空间和动作空间的 MDP 中，该算法运用自模拟度量的方法构建不同任务下状态之间的距离关系，当 2 个 MDP 达到一定相似度时，进行值函数知识迁移，减少算法收敛所需的样本，提高算法的收敛性能；针对强化学习问题中存在的探索与利用的平衡问题，结合贝叶斯推理，利用变分推理获取信息增益并用其构建内部奖赏函数模型，加大 agent 探索力度，提高算法收敛速度。将本文提出的 VFT-HSA

与 Q-Learning 算法、IGP-Sarsa 算法用于经典的 Grid World 问题，实验表明，VFT-HSA 克服了经典的 Sarsa 算法中存在的收敛速度慢以及收敛不稳定的问题，在保证收敛精度的情况下，提高了算法的收敛速度和稳定性。

本文主要在 Grid World 仿真平台中对算法进行实验分析，实验结果表明，本文所提算法具有较快的收敛速度和较好的收敛稳定性。本文主要对较大规模、离散的问题进行实验分析，接下来的工作是将算法运用于更大规模的问题和连续问题中进一步验证算法的有效性。

参考文献：

[1] SUTTON R S, BARTO G A. Reinforcement learning: an introduction[M]. Cambridge: MIT Press, 1998.

[2] SCHMIDHUBER J, INFORMATIK T T. On learning how to learn learning strategies[R]. Germany: Technische University, 1995.

[3] AMMAR H B, EATON E, LUNA J M, et al. Autonomous cross-domain knowledge transfer in lifelong policy gradient reinforcement learning[C]//The 15th International Conference on Artificial Intelligence. 2015: 3345-3351.

[4] GUPTA A, DEVIN C, LIU Y X, et al. Learning invariant feature spaces to transfer skills with reinforcement learning[C]//The 5th International Conference on Learning Representations. 2017: 2147-2153.

[5] LAROCHE R, BARLIER M. Transfer reinforcement learning with shared dynamics[C]//The 31th International Conference on the Association for the Advance of Artificial Intelligence. 2017: 2147-2153.

[6] BARRETO A, DABNEY W, MUNOS R, et al. Successor features for transfer in reinforcement learning[C]//The 32th International Conference on Neural Information Processing Systems. 2017: 4055-4065.

[7] DEARDEN R, NIR F, STUART R. Bayesian Q-learning[C]//The 21th International Conference on the Association for the Advance of Artificial Intelligence. 1998: 761-768.

[8] GUEZ A, SILVER D, DAYAN P. Scalable and efficient Bayes- adaptive reinforcement learning based on Monte-Carlo tree search[J]. Journal of Artificial Intelligence Research, 2013, 48(1): 841-883.

[9] LITTLE D Y, SOMMER F T. Learning and exploration in action-perception loops[J]. Frontiers in Neural Circuits, 2013, 7(7): 37-56.

[10] MANSOUR Y, SLIVKINS A, SYRGKANIS V. Bayesian incentive-compatible bandit exploration[C]//The 16th International Conference on Economics and Computation. 2015: 565-582.

[11] VIEN N A, LEE S G, CHUNG T C. Bayes-adaptive hierarchical MDPs[J]. Applied Intelligence, 2016, 45(1): 112-126.

[12] WU B, FENG Y. Monte-Carlo Bayesian reinforcement learning using a compact factored representation[C]//The 4th International Conference on Information Science and Control Engineering. 2017: 466-469.

[13] 傅启明, 刘全, 伏玉琛, 等. 一种高斯过程的带参近似策略迭代算法[J]. 软件学报, 2013, 24(11): 2676-2687.
FU Q M, LIU Q, FU Y C, et al. Parametric approximation policy strategy iteration algorithm based on Gaussian process[J]. Journal of Software, 2013, 24(11): 2676-2687.

[14] GIVAN R, DEAN T, GREIG M. Equivalence notions and model minimization in Markov decision processes[J]. Artificial Intelligence, 2003, 147(1): 163-223.

[15] FERNS N, PANANGADEN P, PRECUP D. Metrics for finite Markov decision processes[C]//The 20th International Conference on Uncertainty in Artificial Intelligence. 2004: 162-169.

[16] BEAL M J. Variational algorithms for approximate Bayesian inference[D]. London: University of London, 2003.

[17] 傅启明, 刘全, 尤树华, 等. 一种新的基于值函数迁移的快速 Sarsa 算法[J]. 电子学报, 2014, 42(11): 2157-2161.
FU Q M, LIU Q, YOU S H, et al. A novel fast sarsa algorithm based on value function transfer[J]. Acta Electronica Sinica, 2014, 42(11): 2157-2161.

[18] MIERING M, HASSELT H V. The QV family compared to other reinforcement learning algorithms[C]//The 17th International Conference on Approximate Dynamic Programming and Reinforcement Learning. 2008: 101-108.

[19] CHUNG J J, LAWRANCE N R J, SUKKARIEH S. Gaussian processes for informative exploration in reinforcement learning[C]//The 20th International Conference on Robotics and Automation. 2013: 2633-2639.

[作者简介]

陈建平（1963−），男，江苏南京人，博士，苏州科技大学教授，主要研究方向为大数据分析与应用、建筑节能、智能信息处理。

杨正霞（1992−），女，江苏扬州人，苏州科技大学硕士生，主要研究方向为强化学习、迁移学习、建筑节能。

刘全（1969−），男，内蒙古牙克石人，博士，苏州大学教授、博士生导师，主要研究方向为智能信息处理、自动推理与机器学习。

吴宏杰（1977−），男，江苏苏州人，博士，苏州科技大学副教授，主要研究方向为深度学习、模式识别、生物信息。

徐杨（1980−），女，河北深州人，浙江纺织服装职业技术学院讲师，主要研究方向为数据分析与应用、智能化与个性化教学。

傅启明（1985−），男，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、深度学习及建筑节能。