一种新的基于值函数迁移的快速 Sarsa 算法

傅启明[1],刘 全[1,2],尤树华[1],黄 蔚[1],章晓芳[1]
(1.苏州大学计算机科学与技术学院,江苏苏州 215006;2.吉林大学符号计算与知识工程教育部重点实验室,吉林长春 130012)

摘 要: 知识迁移是当前机器学习领域的一个新的研究热点.其基本思想是通过将经验知识从历史任务到目标任务的迁移,达到提高算法收敛速度和收敛精度的目的.针对当前强化学习领域中经典算法收敛速度慢的问题,提出在学习过程中通过迁移值函数信息,减少算法收敛所需要的样本数量,加快算法的收敛速度.基于强化学习中经典的在策略 Sarsa 算法的学习框架,结合值函数迁移方法,优化算法初始值函数的设置,提出一种新的基于值函数迁移的快速 Sarsa 算法——VFT-Sarsa.该算法在执行前期,通过引入自模拟度量方法,在状态空间以及动作空间一致的情况下,对目标任务中的状态与历史任务中的状态之间的距离进行度量,对其中相似并满足一定条件的状态进行值函数迁移,而后再通过学习算法进行学习.将 VTF-Sarsa 算法用于 Random Walk 问题,并与经典的 Sarsa 算法、Q 学习算法以及具有较好收敛速度的 QV 算法进行比较,实验结果表明,该算法在保证收敛精度的基础上,具有更快的收敛速度.

关键词: 强化学习; VFT-Sarsa 算法; 自模拟度量; 值函数迁移
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112 (2014)11-2157-05
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.11.005

A Novel Fast Sarsa Algorithm Based on Value Function Transfer

FU Qi-ming[1], LIU Quan[1,2], YOU Shu-hua[1], HUANG Wei[1], ZHANG Xiao-fang[1]
(1. *Institute of Computer Science and Technology, Soochow University, Suzhou, Jiangsu* 215006, *China*;
2. *Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin* 130012, *China*)

Abstract: Knowledge Transfer has gradually became a research hot pot in machine learning, which tries to transfer the knowledge from the historical tasks to the target task in order to speed up the convergence rate and improve the performance of algorithms. With respect to the slow convergence rate of traditional reinforcement learning algorithms, this paper proposed to transfer the value function between different similar learning tasks with the same state space and action space, which tries to reduce the needed samples in the target task and speed up the convergence rate. Based on the framework of on-policy Sarsa algorithm, combined with the value function transfer method, this paper put forward a novel fast Sarsa algorithm based on the value function transfer—VFT-Sarsa. At the beginning, the algorithm uses Bisimulation metric to measure the distance between states in target task and historical task on the condition that these tasks have the same state space and action space, transfers the value function if the distance meets some condition, and finally executes the learning algorithm. At the end, apply the proposed algorithm in Random Walk, compared with Sarsa algorithm, Q-Learning and QV algorithm, the results show that the proposed algorithm can get a better convergence rate with a good performance.

Key words: reinforcement learning; VFT-Sarsa algorithm; bisimulation metric; value function transfer

1 引言

在强化学习理论研究中,通常同一个算法、同一个 Agent 仅仅针对一个问题,但是在实际应用中,同一个 Agent 在很长的一段时间内通常可能用于处理多个类似的实际问题,例如实际生活中的扫地机器人,它的生命周期可能是 10 年,那么在这 10 年内,它可能用于不同的实际环境中,比如不同的房间、不同的楼层等等,但是它所处理的问题是一样的,比如捡垃圾罐,且这些环境存在一定的共性.因此,可以考虑将以往学习到的经验知识通过某种手段迁移至新的环境中,这就是机器学习领域的知识迁移.通过这种方式,可以使得 Agent 不断重复利用以往学习的经验信息,提高 Agent 解决新问题的速度和精度.

收稿日期:2013-09-12;修回日期:2014-03-03;责任编辑:蓝红杰
基金项目:国家自然科学基金(No.61103045, No.61303108);江苏省自然科学基金(No.BK2012616);江苏省高校自然科学研究项目(No.13KJB520020);吉林大学符号计算与知识工程教育部重点实验室资助(No.93K17012K04)

目前,关于知识迁移在强化学习领域的研究已经引起研究者的广泛关注.针对马尔科夫决策问题(Markov Decision Process,MDP),国外的 Taylor 等人以及国内的王皓等人都对当前强化学习领域现存的知识迁移方法进行了全面的介绍[1,2];Sunmola 和 Wyatt 通过参数匹配方法将目标任务的先验信息与历史任务的经验信息进行关联,构造更为精确的任务模型,加速算法的收敛[3];Konidaris 和 Barto 利用 MDP 之间的同态特征以及构造合理的 option,实现具有不同状态空间和动作空间的问题之间的知识迁移[4];Ferrante 在假设不同任务共享相同的状态空间和动作空间的基础上,通过构造基于策略的原型值函数,实现不同任务之间的信息迁移,加快算法的收敛[5];Lasaric 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和奖赏函数的情况下,通过将历史任务中收集到的样本数据迁移到目标任务中,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,减少在目标任务中所需要的样本数据,加快算法的收敛[6];Sorg 和 Singh 提出利用 MDP 之间的软同态特征,构造不同任务中状态之间的映射关系,实现不同任务间的迁移学习,并给出由知识迁移所导致的损失函数的理论边界[7];Ammar 等人通过构造不同任务之间的映射关系,实现类似任务之间信息的迁移,减少算法在后续任务中收敛所需要的样本数量,提高算法的收敛性能[8];Konidaris 等人在假设任务间存在类似特征空间的情况下,通过共享特征实现不同任务之间信息的迁移,提高算法在后续任务中的收敛性能[9].

对于迭代强化学习算法来讲,值函数初始值的设置直接影响算法的收敛速度,考虑最极端的情况,值函数的初始值就是值函数的最优值,则算法只需要极少量的样本就发现值函数已经趋于稳定.强化学习中的"乐观"初始值就是在对问题有一定先验知识的情况下,设置合理的值函数,鼓励探索,加速算法的收敛[10].本文主要从值函数的角度实现知识的迁移,提出一种新的基于值函数迁移的快速 Sarsa 算法——VFT-Sarsa (Sarsa Algorithm Based-on Value Function Transfer).算法利用自模拟度量方法,在两个任务的状态空间以及动作空间一致的情况下,构造目标任务中状态和历史任务中状态之间的度量关系,当其满足一定的阈值时,对值函数进行迁移.值得注意的是只有当两个状态之间的距离满足一定阈值时,值函数的迁移才能对学习有促进作用,可以称作"正迁移",而不满足该阈值关系,则容易阻碍学习的收敛,可以称作"负迁移".因此,算法通过自模拟度量,构造状态之间的度量关系,尽量实现知识的"正迁移",而后结合学习算法再与环境交互学习,加速学习算法的收敛.

2 马尔科夫决策过程

马尔科夫决策过程可以用来对强化学习问题进行建模,通常定义为一个四元组,$M=(X,U,\rho,f)$,其中 X 是状态集合;U 是动作集合;ρ 是奖赏函数,$\rho: X\times U\to\mathbb{R}$,例如 $\rho(x,u)$表示在状态 x 下采用动作 u 所获得的立即奖赏;f 是状态转移函数,$f: X\times U\times X\to[0,1]$,例如 $f(x,u,x')$表示在状态 x 下采用动作 u 转移到状态 x'的概率.

强化学习的最终目标是要能够学习到一个策略,利用该策略进行决策.假设在时刻 k,状态为 x_k,策略为 h,Agent 根据当前状态 x_k 以及策略 h 选择动作 u_k,获得立即奖赏为 $\rho(x_k,u_k)$,状态转移至 x_{k+1},在算法的学习过程中,不断重复该过程,直至获得最优策略 h^*.为了衡量策略 h 的优劣,强化学习中引入值函数的概念,利用值函数评估策略,具体分为状态值函数 $V^h(x)$和动作值函数 $Q^h(x,u)$,如公式(1)和(2)所示.

$$V^h(x)=\sum_{u\in U}h(x,u)[\rho(x,u)+\gamma\sum_{x'\in X}f(x,u,x')V^h(x')] \tag{1}$$

$$Q^h(x,u)=\rho(x,u)+\gamma\sum_{x'\in X}f(x,u,x')\sum_{u'\in U}h(x',u')Q^h(x',u') \tag{2}$$

其中 γ 是折扣因子.

3 自模拟度量与状态之间的距离

自模拟关系由 Givan 等人在 2003 年首次引入 MDP,并度量 MDP 中状态之间的关系[11].简单来讲,如果两个状态满足自模拟关系,则这两个状态应共享相同的最优值函数及最优动作.

定义 1 自模拟(Bisumulation)关系.若关系 $E\subseteq X\times X$ 是自模拟关系,则对于 $x',x''\in X$,$x'Ex''$满足以下性质:(1)对于 $\forall u\in U$,$\rho(x',u)=\rho(x'',u)$;(2)对于 $\forall u\in U$,$\forall C\in X/E$,$\sum_{t\in C}f(x',u,t)=\sum_{t\in C}f(x'',u,t)$,其中$X/E$是状态集合 X 关于 E 的等价集合.若 x',$x''\in X$,满足自模拟关系,可记作:$x'\sim x''$.

对于任意两个状态,两者之间的自模拟关系是"是"或者"非"的关系,即要么两者满足自模拟关系,要么两者不满足自模拟关系.但是,在实际应用中,只有在两个状态的奖赏分布和状态转移概率分布完全一致的情况下,才认为这两个状态是满足自模拟关系的.然而,如果两个状态的奖赏分布以及状态转移概率分布仅存在很小的差别,在状态空间中,这两个状态是非常接近的,但是自模拟关系却无法区分这种微小的差别.针对这个问题,Ferns 等人在自模拟关系的基础之上,利用 Kantorovich 距离衡量两个概率分布之间的距离,提出

一种可用于衡量两个状态之间远近关系的自模拟度量方法(Bisimulation Metric)[12].

定理 1　D 为定义在状态集 X 上的度量集合,且度量 $d\in D$.对于 $\forall x', x''\in X$,定义 $G: D\to D$, $G(d)(x', x'') = \max_{u\in U}(d_u(x', x'') + \gamma T_K(d)(f(x', u, \cdot), f(x'', u, \cdot))$,其中 $d_u(x', x'') = |\rho(x', u) - \rho(x'', u)|$, $0<\gamma<1$,则 G 存在一个不动点 $d_\sim$,且 $d_\sim$ 是自模拟度量,$d_\sim(x', x'')$是状态 x' 和 x'' 之间的距离.

关于该定理的证明请参考 Frens 等人的论文[12].同时,Frens 等人还证明,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 $\left\lceil \frac{\ln\zeta}{\ln\gamma} \right\rceil$.

4　基于值函数迁移的 Sarsa 算法

4.1　基于自模拟度量的值函数迁移

假设 1　两个 MDP, M_1 和 M_2,具有相同的离散状态集合 X 以及离散动作集合 U,不同的奖赏函数和状态转移函数,即 $M_1 = \langle X, U, f_1, \rho_1\rangle$, $M_2 = \langle X, U, f_2, \rho_2\rangle$.

满足假设 1 的两个 MDP, M_1 和 M_2,其中 M_1 是原始 MDP(用于迁移值函数的 MDP), M_2 是目标 MDP(被迁移值函数的 MDP).令 V_1^* 和 h_1^* 分别是 M_1 的最优状态值函数和最优策略, V_2^* 和 h_2^* 分别是 M_2 的最优状态值函数和最优策略.为了区分两个状态来自不同的 MDP,分别表示为 x_1 和 x_2,其后续状态分别为 x_1' 和 x_2',其中 x_1 和 x_1' 是 M_1 中的状态, x_2 和 x_2' 是 M_2 中的状态.

定理 2　假设 $d_\sim(x_1, x_2) = \delta$,则 $|V_1^*(x_1) - V_2^*(x_2)|\leqslant\delta$.

证明:根据最优状态值函数的定义,对 $|V_1^*(x_1) - V_2^*(x_2)|$ 展开可得,

$$
\begin{aligned}
&|V_1^*(x_1) - V_2^*(x_2)|\\
&= \Big|\max_{u_1\in U}\{\rho(x_1, u_1) + \gamma\textstyle\sum_{x_1'\in X} f_1(x_1, u_1, x_1')V_1^*(x_1')\}\\
&\quad - \max_{u_2\in U}\{\rho(x_2, u_2) + \gamma\textstyle\sum_{x_2'\in X} f_2(x_2, u_2, x_2')\\
&\quad V_2^*(x_2')\}\Big|\\
&\leqslant \max_{u\in U}\big|(\rho(x_1, u) + \gamma\textstyle\sum_{x_1'\in X} f_1(x_1, u, x_1')V_1^*(x_1'))\\
&\quad - (\rho(x_2, u) + \gamma\textstyle\sum_{x_2'\in X} f_2(x_2, u, x_2')V_2^*(x_2'))\big|\\
&= \max_{u\in U}|d_u(x_1, x_2) + \gamma T_K(d)(f_1(x_1, u, \cdot),\\
&\quad f_2(x_2, u, \cdot))|\\
&= d_\sim(x_1, x_2)\\
&= \delta
\end{aligned}
$$

因此,对于 $\forall x_1, x_2\in X$, $|V_1^*(x_1) - V_2^*(x_2)|\leqslant d_\sim(x_1, x_2) = \delta$ 成立.　证毕.

定义 2　**正迁移**.满足假设 1 的两个 MDP, M_1 和 M_2,为了区别两者的状态空间, X_1 表示 M_1 的状态空间, X_2 表示 M_2 的状态空间, V_1^* 是 M_1 的最优值函数.给定阈值 ξ,对于 $\forall x_2\in X_2$,

$$
V_2(x_2) = \begin{cases} V_1^*(x), & \text{if } x = \operatorname{argmin}_{x\in X_1} d_\sim(x, x_2) \text{ and } d_\sim(x, x_2)\leqslant\xi\\ 0, & \text{other}\end{cases}
$$

当 $x = \operatorname{argmin}_{x\in X_1} d_\sim(x, x_2)$且 $d_\sim(x, x_2)\leqslant\xi$ 时,所进行的值函数迁移被称作"正迁移".

算法 1　基于自模拟度量的值函数迁移算法

1:输入:两个 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 ξ
2:For $k = 1$ to $k <= |X_1|$ do
3:　For $m = 1$ to $m <= |X_2|$ do
4:　　计算 $d_\sim(x_k, x_m)$
5:　End For
6:End For
7:For $i = 1$ to $i <= |X_2|$ doc
8:　$x = \operatorname{argmin}_{x\in X_1} d_\sim(x, x_i)$
9:　If $d_\sim(x, x_i)\leqslant\xi$ then $V_2(x_i) = V_1^*(x)$
10:　Else $V_2(x_i) = 0$
11:　End If
12:End for
13:输出: V_2

4.2　VFT-Sarsa

基于值函数迁移的 VFT-Sarsa 算法主要利用基于自模拟度量的值函数迁移方法,对历史值函数信息在相似状态之间迁移,结合 Sarsa 算法的框架,利用 *Q-V* 算法中状态值函数及动作值函数的更新方法更新值函数[13],加快算法收敛.

算法 2　VFT-Sarsa 算法

1:输入:阈值参数 δ,学习因子 α, β,贪心因子 ε
2:初始化:利用算法 2 初始化状态值函数 V_0,且对于 $\forall(x, u)\in X\times U$, $Q_0(x, u) = 0$
3:令 $k = 1$
4:Repeat(对于每一个情节)
5:　初始状态动作对 (x, u)
6:　Repeat(对于情节中的每一个时间步)
7:　　x' 是 x 下采用动作 u 的后续动作,奖赏值为 r,并利用 ε - *greedy* 策略选择 x' 下的动作 u'
8:　　$V_k(x) = V_{k-1}(x) + \alpha(r + \gamma V_{k-1}(x') - V_{k-1}(x))$
9:　　$Q_k(x, u) = Q_{k-1}(x, u) + \beta(r + \gamma V_{k-1}(x') - Q_{k-1}(x, u))$
10:　　令 $x\leftarrow x'$,且 $u\leftarrow u'$
11:　End Repeat
12:　If $\| Q_k - Q_{k-1}\|_\infty\leqslant\delta$ then 算法终止
13:　End If

14: $k = k + 1$

15: End Repeat

16: 输出: 对于 $\forall x \in X$, 策略 $h(x) = \underset{u \in U}{\operatorname{argmax}} Q(x, u)$

5 实验结果分析

5.1 问题描述

Random Walk 是一个包含 7 个状态的 Markov 链, 如图 1 所示, 链的两端是两个吸收状态, 状态 x_3 是每个情节的初始状态, 到达吸收状态则情节结束. 在任意状态下, Agent 有两个动作可供选择——u_0 和 u_1, 其中 u_0 和 u_1 分别表示向右、向左的动作. 在该问题中, 当 Agent 选择任意动作时, 有一定的概率执行当前所选择的动作并进行状态的迁移, 也有一定的概率执行相反的动作并进行相应的状态迁移, 即动作的选择存在"随机滑动". 在状态迁移过程中, 到达状态 x_6, 相应的立即奖赏是 5, 其他情况, Agent 所获得的立即奖赏是 1.

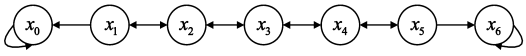

图1 Random Walk示意图

5.2 实验分析

图 2 和图 3 主要用于说明 VFT-Sarsa、Q-Learning、Sarsa 以及 QV 算法[13]在具有随机性的 Random Walk 问题上平均性能(在实验过程中, 每个算法都被独立执行 20 次). 从图 2 可以看出, VFT-Sarsa 的收敛速度要明显优于 Q-Learning、Sarsa 以及 QV 算法, 这主要是由于通过值函数迁移, 提高初始值函数的精确性, 加速了算法的收敛; 同时, 在收敛精度上, VFT-Sarsa 要优于 Q-Learning 和 Sarsa, 且与 QV 算法类似, 这主要是由于 VFT-Sarsa 中采用了类似 QV 算法的 Q 值函数和 V 值函数的更新规则, 而通过值函数迁移, 仅仅使 VFT-Sarsa 中初始 V 值函数更加接近最优 V 值函数, 减少算法收敛所需要的迭代次数, 加快算法的收敛, 但是通过不断迭代, 两个算法最终具有类似的值函数.

图 3 是以状态 x_3 为分析目标, 说明在状态 x_3 下最优动作的选择情况. 实验代码中用 0 表示动作 u_0, 用 1 表示 u_1, 因此, 统计数据中各情节下动作值比例越靠近 0, 则说明当前选择最优动作 u_0 的概率越高. 从图 3 可以看出, 针对状态 x_3, VFT-Sarsa 收敛至最优动作 u_0 的速度要明显优于 Q-Learning、Sarsa 以及 QV 算法. 另外值得注意的是, 图 3 中四个算法都没有稳定收敛至动作 u_0, 这主要是由于在整个学习过程中, 都是采用 ε – greedy 算法选择动作, 即使 Q 值函数收敛之后, 在动作的选择上依然存在一定的概率选择次优动作, 但这并不影响算法已经收敛至最优动作(因为根据 Q 值函数, 可以很轻易地得出最优动作). 因此, 综合以上分析, 发现与 Q-Learning, Sarsa 以及 QV 算法相比, VFT-Sarsa 在保证收敛精度的基础之上, 具有较快的收敛速度, 同时也证明值函数迁移的有效性.

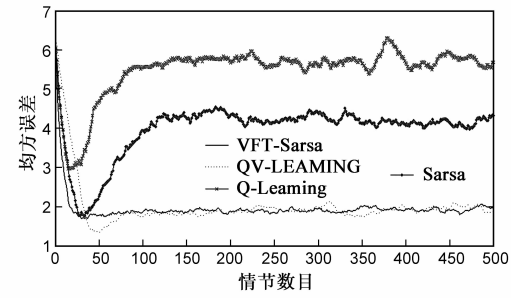

图2 各算法在不同情节数下Q值函数均方差

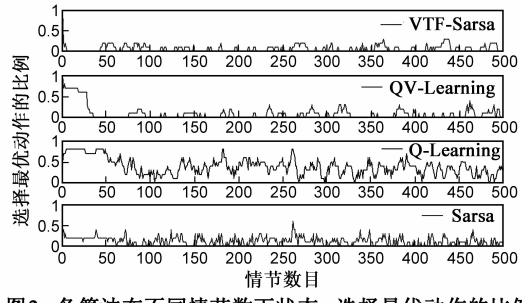

图3 各算法在不同情节数下状态x_3选择最优动作的比例

表 1 说明两个 MDP 最优值函数之间的误差与状态之间距离的关系, 即用于验证定理 2 的正确性. 在实验的过程中, 由于两个 MDP 设置的特殊性——仅仅动作的滑动概率不一致, 因此, 相应的状态之间的距离是最小的, 即与第二个 MDP 中的状态 x_2 距离最小的状态是第一个 MDP 中的对应的状态 x_2. 但是这种情况并不是绝对的, 在其他复杂问题中, 可能会出现不对应的情况, 即与第二个 MDP 中 x_2 对应的是第一个 MDP 中的 x_1, 但是这并不影响值函数的迁移, 只要所迁移的值函数的值更加接近被迁移的状态的最优值函数的值, 而这恰恰又是被定理 2 所保证的. 根据表 1 可以看出, 在该问题中, 对应状态之间最优值函数小于对应状态之间的距离, 这也从实验的角度验证定理 2 的正确性.

表 1 最优值函数误差与状态之间的距离

状态	最优值函数误差	对应状态的距离
x_1	2.07051	2.288379
x_2	1.448669	1.693082
x_3	1.107099	1.139326
x_4	0.823076	1.161157
x_5	0.477113	0.809682

6 结束语

本文从知识迁移的角度, 在不同任务之间具有相

同状态空间和动作空间的情况下,利用自模拟度量方法构造不同环境下状态之间的距离关系,设置一定的阈值,根据状态之间的距离,实现不同环境下状态值函数之间的"正迁移",并从理论上证明不同环境下状态最优值函数之间的误差与状态之间距离的关系,理论上保证值函数迁移的正确性.基于值函数迁移算法,利用 Sarsa 算法的执行框架,结合 QV 算法中值函数更新规则,提出一种基于值函数迁移的 VFT-Sarsa 算法.将 VFT-Sarsa、Q-Learning、Sarsa 以及 QV 算法用于带有随机性的 Random Walk 实验平台,实验结果表明,基于自模拟度量的值函数迁移的正确性以及 VFT-Sarsa 在保证收敛精度的情况下,具有较快的收敛速度.

本文主要针对基于查询表的值函数进行不同环境下的值函数迁移,但是在大规模状态空间或者连续状态空间中,却难以利用查询表表示值函数空间,且进行相应的值函数迁移.因此,未来的工作就是考虑如何将值函数迁移方法与基于函数近似的强化学习方法相结合,构造不同环境下特征空间的对应关系,进行近似值函数的迁移,加快算法的收敛.

参考文献

[1] Taylor J, Precup D, Panangaden P. Bounding performance loss in approximate MDP homomorphisms[A]. Proceedings of the 22nd Annual Conference on Neural Information Processing Systems[C]. NY: Curran Associates, 2008. 1660-1667.

[2] 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(12A):39 – 43.
Wang Hao, Gao Yang, Chen Xinguo. Transfer of reinforcement learning: The state of the art[J]. Acta Electronica Sinica, 2008, 36(12A): 39 – 43. (in Chinese)

[3] Sunmola F T, Wyatt J L. Model transfer for Markov decision tasks via parameter matching[A]. Proceedings of the 25th Workshop of the UK Planning and Scheduling Special Interest Group[C]. Nottingham, England, 2006. 17 – 24.

[4] Konidaris G D, Barto A G. Building portable options: skill transfer in reinforcement learning[A]. Proceedings of the 20th International Joint Conference on Artificial Intelligence[C]. CA: Morgan Kaufmann Publishers, 2007. 895 – 901.

[5] Ferrante E, Lazaric A, Restelli M. Transfer of task representation in reinforcement learning using policy-based proto-value functions[A]. Proceedings of the 7th International Conference on Autonomous Agents and Multi-Agent Systems[C]. Estoril:, 2008. 1329 – 1332.

[6] Lazaric A, Restelli M, Bonarini A. Transfer of samples in batch reinforcement learning[A]. Proceedings of the 25th International Conference on Machine Learning[C]. NY: ACM Press, 2008. 544 – 551.

[7] Sorg J, Singh S. Transfer via soft homomorphisms[A]. Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems[C]. Hungary:, 2009. 741 – 748.

[8] Ammar H B, Taylor M, Tuyls K, Weiss G. Reinforcement learning transfer using a sparse coded inter-task mapping[A]. Proceedings of the 9th European Workshop on Multi-agent Systems[C]. Berlin: Springer-verlag, 2012. 1 – 16.

[9] Konidaris G D, Scheidwasser I and Barto A G. Transfer in Reinforcement Learning via Shared Features[J]. Journal of Machine Learning Research, 2012, 13: 1333 – 1371.

[10] Sutton R S, Barto A G. Reinforcement Learning[M]. Cambridge: MIT Press, 1998.

[11] Givan R, Dean T, Greig M. Equivalence notions and model minimization in Markov decision processes[J]. Artificial Intelligence, 2003, 147(1 – 2): 163 – 223.

[12] Ferns N, Panangaden P, Precup D. Metrics for finite markov decision processes[A]. Proceeding of the 20th Conference on Uncertainty in Artificial Intelligence[C]. Arlington: AUAI Press, 2004. 162 – 169.

[13] Wiering M andHasselt H V. The QV family compared to other reinforcement learning algorithms[A]. Proceedings of IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning[C]. Nashville: IEEE, 2009. 101 – 108.

作者简介

傅启明　男,1985 年生于江苏淮安,博士生.主要研究方向为强化学习、贝叶斯推理及遗传算法.

刘　全(通信作者)　男,1969 年生于内蒙古,博士,教授,博士生导师.主要研究方向为强化学习、无线传感器网络、智能信息处理.
E-mail: quanliu@suda.edu.cn