一类基于谱方法的强化学习混合迁移算法

朱美强 [1] 程玉虎 [1] 李 明 [1] 王雪松 [1] 冯涣婷 [1]

摘 要 在状态空间比例放大的迁移任务中, 原型值函数方法只能有效迁移较小特征值对应的基函数, 用于目标任务的值函数逼近时会使部分状态的值函数出现错误. 针对该问题, 利用拉普拉斯特征映射能保持状态空间局部拓扑结构不变的特点, 对基于谱图理论的层次分解技术进行了改进, 提出一种基函数与子任务最优策略相结合的混合迁移方法. 首先, 在源任务中利用谱方法求取基函数, 再采用线性插值技术将其扩展为目标任务的基函数; 然后, 用插值得到的次级基函数 (目标任务的近似 Fiedler 特征向量) 实现任务分解, 并借助改进的层次分解技术求取相关子任务的最优策略; 最后, 将扩展的基函数和获取的子任务策略一起用于目标任务学习中. 所提的混合迁移方法可直接确定目标任务部分状态空间的最优策略, 减少了值函数逼近所需的最少基函数数目, 降低了策略迭代次数, 适用于状态空间比例放大且具有层次结构的迁移任务. 格子世界的仿真结果验证了新方法的有效性.

关键词 强化学习, 迁移学习, 谱图理论, 原型值函数, 层次分解

引用格式 朱美强, 程玉虎, 李明, 王雪松, 冯涣婷. 一类基于谱方法的强化学习混合迁移算法. 自动化学报, 2012, **38**(11): 1765−1776

DOI 10.3724/SP.J.1004.2012.01765

A Hybrid Transfer Algorithm for Reinforcement Learning Based on Spectral Method

ZHU Mei-Qiang [1] CHENG Yu-Hu[1] LI Ming[1] WANG Xue-Song[1] FENG Huan-Ting[1]

Abstract For scaling up state space transfer underlying the proto-value function framework, only some basis functions corresponding to smaller eigenvalues are transferred effectively, which will result in wrong approximation of value function in the target task. In order to solve the problem, according to the fact that Laplacian eigenmap can preserve the local topology structure of state space, an improved hierarchical decomposition algorithm based on the spectral graph theory is proposed and a hybrid transfer method integrating basis function transfer with subtask optimal polices transfer is designed. At first, the basis functions of the source task are constructed using spectral method. The basis functions of target task are produced through linearly interpolating basis functions of the source task. Secondly, the produced second basis function of the target task (approximating Fiedler eigenvector) is used to decompose the target task. Then the optimal polices of subtasks are obtained using the improved hierarchical decomposition algorithm. At last, the obtained basis functions and optimal subtask polices are transferred to the target task. The proposed hybrid transfer method can directly get optimal policies of some states, reduce the number of iterations and the minimum number of basis functions needed to approximate the value function. The method is suitable for scaling up state space transfer task with hierarchical control structure. Simulation results of grid world have verified the validity of the proposed hybrid transfer method.

Key words Reinforcement learning, transfer learning, spectral graph theory, proto-value functions, hierarchical decomposition

Citation Zhu Mei-Qiang, Cheng Yu-Hu, Li Ming, Wang Xue-Song, Feng Huan-Ting. A hybrid transfer algorithm for reinforcement learning based on spectral method. *Acta Automatica Sinica*, 2012, **38**(11): 1765−1776

强化学习作为一类求解序贯优化决策问题的有效方法, 已在自动控制、运筹学和计算科学等领域得到了广泛应用[1−3]. 强化学习的最大特点是能在无环境模型、无教师样本的情况下, 通过与环境交互试错来极大化累计回报, 从而获得最优或者次优行为策略. 在强化学习问题中, 学习算法需要对状态–动作序列进行足够多次的访问才能收敛, 学习效率较低, 阻碍了它在大规模或连续问题中的应用[1].

收稿日期 2011-12-02 录用日期 2012-05-22
Manuscript received December 2, 2011; accepted May 22, 2012
国家自然科学基金 (60974050, 61072094, 61273143), 中国矿业大学青年科技基金 (OC080252), 教育部新世纪优秀人才支持计划 (NCET-08-0836, NCET-10-0765), 教育部高等学校博士学科点专项科研基金 (20110095110016) 资助
Supported by National Natural Science Foundation of China (60974050, 61072094, 61273143), Youth Science and Technology Foundation of China University of Mining and Technology (OC080252), Program for New Century Excellent Talents in University (NCET-08-0836, NCET-10-0765), and Specialized Research Fund for the Doctoral Program of Higher Education of China (20110095110016)
本文责任编委 陈杰
Recommended by Associate Editor CHEN Jie

1. 中国矿业大学信息与电气工程学院 徐州 221116
1. School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou 221116

解决复杂问题中强化学习效率低的常用方法有值函数逼近、分层强化学习、迁移学习和状态空间抽象等[2]. 值函数逼近使用函数逼近器替代传统强化学习中的查表法来实现泛化[1]. 分层强化学习采用分而治之原则, 把一个复杂问题分解成多个简单问题来求解[4]. 迁移学习则是模拟人举一反三、触类旁通的能力, 根据不同学习任务间的相互联系, 利用过去的学习经验来加速新任务的学习[5−7]. 上述方法分别从不同角度出发, 抛弃过去研究中智能体 (Agent) "一无所知" 的假设, 通过发现和利用问题的领域知识来提高学习效率, 多种方法可以结合使用. Mahadevan 等提出的原型值函数 (Proto-value functions, PVF) 方法[8] 不仅能产生便于在相似任务间迁移的正交基函数, 还可实现任务分解, 为分层强化学习、值函数逼近和迁移学习的结合提供了一种框架.

在值函数逼近方法中, 逼近器的结构和参数选取直接影响算法的学习效率, 已有研究多是人为手工设计的. 基于谱图理论的 PVF 方法利用问题的内在结构自动产生反映任务全局光滑性的正交基函数, 其中最光滑的次级基函数 (Second proto-value function, SPVF), 即 Fiedler 特征向量, 能用于任务分解[9−10]. PVF 方法是一类抽象的傅里叶变换, 其小特征值形成的基函数 (称为 "低频" 部分) 决定了被逼近函数的重要特征[11]. 若两个强化学习问题的状态连接图的结构相似, 则其 SPVF 隐含的任务层次结构和光滑的 "低频" 基函数也类似, 两者均可用于迁移学习. 已有的 PVF 迁移研究只考虑了基函数的迁移, 忽视了相似层次结构信息的利用. 文献 [11] 讨论了在状态连接图不变、微小变化和比例缩放三种情况下 PVF 基函数迁移学习的效果. 其实验结果表明: 对于状态连接图保持不变和变化很小的任务, PVF 基函数迁移具有较好效果; 在状态空间比例放大的任务中, 采用线性插值技术的迁移效果不稳定. 本文通过分析得到迁移效果不稳定的主要原因: 插值所获得的有效的基函数数目过少致使逼近能力不足. 然后, 通过合理利用任务的层次结构信息来解决此问题.

在基于谱方法的层次强化学习中, 已有研究通常借助 SPVF 来寻找瓶颈状态, 以图分割的方式实现任务分解. 在完成任务分解后, 子任务中相关策略的求取一般都作为一个新的强化学习问题来处理[9−10]. 事实上, 由拉普拉斯特征映射得到的每个 PVF 都是状态空间至一维实数空间的一种映射, n 个最光滑的低频 PVF 的组合可看作是状态空间至 n 维实数空间的映射. 这类映射得到的像很好地保留了原有状态空间的拓扑结构: 其图上距离很近的状态所对应的像间的欧氏距离也很近. 这个特点常被用于谱聚类和流形学习[12−13], 本文将它应用到子任务的策略求取中, 用于改进基于谱图理论的层次分解技术.

针对状态空间具有层次结构的强化学习任务, 本文改进了基于谱图分割的层次分解技术. 在此基础上, 将基函数迁移和层次结构信息迁移相结合, 提出一种基于谱方法的强化学习混合迁移方法. 对于状态连接图变化较小的任务, 迁移方法直接将源任务得到的基函数和子任务的最优策略用于目标任务的学习中. 对于状态空间比例放大的情况, 则首先使用线性插值技术将源任务的基函数扩展为目标任务的 "低频" 基函数, 然后再利用插值所得的近似 SPVF 实现目标任务的层次分解. 最后, 将扩展的基函数和获取的子任务策略一起用于目标任务的学习中.

1 背景知识

1.1 最小二乘策略迭代

强化学习以马尔科夫决策过程 (Markov decision process, MDP) 为基础, 一个离散的 MDP 模型可表示为: $\{S, A, \mathcal{P}, r, \gamma\}$. 其中, S 和 A 分别为有限状态空间和动作空间, $\mathcal{P}(s, a, s')$ 和 $r(s, a, s')$ 分别表示状态 s 执行动作 a 转移到状态 s' 的概率和得到的回报. $\gamma \in (0, 1]$ 为折扣因子, 状态–动作对 (s, a) 的期望回报 $\mathcal{R}(s, a)$ 为

$$\mathcal{R}(s, a) = \sum_{s' \in S} \mathcal{P}(s, a, s') r(s, a, s') \tag{1}$$

设确定性策略 $\pi(s) : S \to A$ 是从状态到动作的映射, 其动作值函数 $Q^{\pi}(s, a) : S \times A \to \mathbf{R}$ 是状态–动作对到实数的映射, 表示智能体在状态 s 选择动作 a 后, 按照策略 π 执行获得的长期折扣回报的期望. 根据贝尔曼公式有[1]:

$$Q^{\pi}(s, a) = \mathcal{R}(s, a) + \gamma \sum_{s' \in S} \mathcal{P}(s, a, s') Q^{\pi}(s', \pi(s')) \tag{2}$$

强化学习的目标是在转移模型 $\mathcal{P}$ 和期望回报 $\mathcal{R}$ 未知的情况下学习一个可最大化长期累计折扣回报的最优策略 π^* :

$$\pi^*(s) = \arg\max_{a \in A} Q^*(s, a) \tag{3}$$

其中, $Q^*(s, a) = \max_{\pi} Q^{\pi}(s, a)$ 为最优动作值函数. 求取最优策略最常用的方法是策略迭代, 策略迭代分为策略评估和策略改进两部分:

$$\pi_1 \xrightarrow{E} Q^{\pi_1} \xrightarrow{I} \pi_2 \xrightarrow{E} Q^{\pi_2} \xrightarrow{I} \cdots \xrightarrow{I} \pi^* \xrightarrow{E} Q^* \tag{4}$$

其中, π_1 为初始策略, E 表示策略评估, 即用于计算值函数 $Q^{\pi}(s, a)$. I 表示策略改进, 常用式 (5) 所示的贪婪更新方式. 由于模型未知, 策略评估中不能直

接计算 $Q^{\pi}(s,a)$, 所以需要根据蒙特卡洛或时间差分学习方法对其估计.

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

函数逼近法是解决强化学习维数灾难的一种有效方法, 常用的函数逼近器主要有非线性结构的神经网络、参数化的线性函数逼近器和非参数化的支持向量机等[14−16]. 其中, 线性函数逼近器应用更为广泛, 最小二乘策略迭代 (Least squares policy iteration, LSPI) 就是一种将线性逼近器与最小二乘时间差分算法相结合的策略迭代方法[14]. LSPI 常用的值函数逼近方法有不动点法和贝尔曼残差最小化法, 下面以不动点法为例来介绍其原理.

引入线性函数逼近器后, 动作值函数 $Q^{\pi}(s,a)$ 可表示为

$$\hat{Q}^{\pi}(s,a)=\sum_{i=1}^{d}\varphi_i(s,a)w_i \tag{6}$$

其中, $\boldsymbol{w}^{\pi}=(w_1,w_2,\cdots,w_d)^{\mathrm{T}}$ 为模型参数, $\boldsymbol{\varphi}(s,a)=(\varphi_1(s,a),\varphi_2(s,a),\cdots,\varphi_d(s,a))$ 为预先选定的基函数向量, d 为基函数向量的长度 (基函数的个数), 且 $d\ll|S|\times|A|$. 式 (6) 的矩阵形式可表示为

$$\hat{Q}^{\pi}=\Phi\boldsymbol{w}^{\pi} \tag{7}$$

$\Phi=\left[\boldsymbol{\varphi}(s_1,a_1),\cdots,\boldsymbol{\varphi}(s,a),\cdots,\boldsymbol{\varphi}\left(s_{|S|},a_{|A|}\right)\right]^{\mathrm{T}}$ 表示 $|S||A|\times d$ 的基函数矩阵. 根据最小二乘不动点原理, 利用式 (8) 可以近似求解权值 $\boldsymbol{w}^{\pi}$[14].

$$\boldsymbol{w}^{\pi}=\left[\Phi^{\mathrm{T}}\left(I-\gamma P\Pi_{\pi}\right)\Phi\right]^{-1}\Phi^{\mathrm{T}}\mathcal{R} \tag{8}$$

其中, P 为 $|S||A|\times|S|$ 的矩阵, $P((s,a),s')=\mathcal{P}(s,a,s')$, Π_{π} 为 $|S|\times|S||A|$ 的矩阵, $\Pi_{\pi}(s',(s',a'))=\pi(s')$. 由于问题模型未知, 式 (8) 不能直接计算, 可根据样本数据进行迭代求解. 假设在任意策略 π 下得到 N 个样本 $\{(s_t,a_t,s'_t,r_t)\}_{t=1}^{N}$, 则有 $\boldsymbol{w}^{\pi}=\tilde{A}^{-1}\boldsymbol{b}$, 其中:

$$\begin{cases}\tilde{A}=\sum\limits_{t=1}^{N}\boldsymbol{\varphi}(s_t,a_t)\left(\boldsymbol{\varphi}(s_t,a_t)-\gamma\boldsymbol{\varphi}(s'_t,\pi(s'_t))\right)^{\mathrm{T}}\\ \boldsymbol{b}=\sum\limits_{t=1}^{N}\boldsymbol{\varphi}(s_t,a_t)\,r_t\end{cases} \tag{9}$$

1.2　强化学习中的迁移

人有举一反三的能力, 我们自然希望学习算法也如此, 能借用已解决问题的相关知识来辅助相近任务的学习, 从而引出了知识迁移 (Knowledge transfer) 的研究[5]. 知识迁移作为机器学习领域的一项新兴技术, 近年来在监督学习中广受关注, 最近被引入到强化学习中[6−7].

引入迁移学习是提高强化学习效率的一条有效途径. 迁移学习的任务分为源任务和目标任务, 任务间的迁移主要存在三个研究主题: 迁移什么、如何迁移与何时迁移. 在强化学习中, 源任务中的采样样本、回报函数、值函数初值、值函数逼近器结构和参数、模型、策略、建议规则、问题背景知识、任务层次结构等都可以迁移到目标任务中 (见图 1). 具体迁移什么、怎样迁移, 取决于任务相似程度和特点[6]. 通常, 寻找任务间的公共的子任务 (Option 方法)、不变的环境拓扑和高层的领域知识进行迁移较为有效[7].

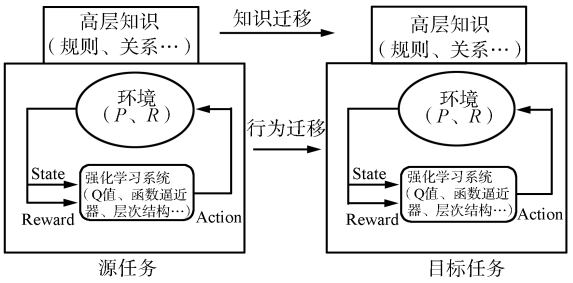


图 1　强化学习中的迁移方法

Fig. 1　Transfer method for reinforcement learning

迁移学习首先要解决任务相似度度量的问题. 通常, 源任务和目标任务的相似度可分为三种: 源样本与目标样本的相似度、源样本与目标任务的相似度、源任务与目标任务的相似度. 对于实际问题, 样本与任务的相似度偏差较大, 任务相似度的度量才是关键. 文献 [6] 从马尔科夫模型 $\{S,A,\mathcal{P},r,\gamma\}$ 的相似度入手, 把强化学习迁移分为三类: 状态和动作变量维数相同的任务迁移、多任务迁移和状态与动作变量维数不同的任务迁移. 基于谱图理论的 PVF 方法可以提取基于环境拓扑的基函数并实现子任务分解, 能将层次强化学习和值函数泛化方法结合起来迁移, 对于具有层次结构的第一类任务有较好的研究前景.

1.3　原型值函数

由式 (6) 可知, 线性基函数的选择对动作值函数的计算影响很大, 直接决定强化学习算法的效率. 常用的基函数有径向基函数、多项式基函数和傅里叶级数等, 它们通常需要人工设定相关参数, 不能有效地逼近欧氏空间内不连续的值函数 (例如迷宫问题)[8]. 而基于谱图理论的 PVF 基函数则可以较好地解决这些问题, 并能实现任务分解.

谱图理论的主要思想是通过分析图上定义的拉普拉斯矩阵的谱和相应的特征向量来揭示图所包含的信息[17]. 常用的图拉普拉斯矩阵有组合拉普拉斯矩阵、正则化拉普拉斯矩阵和随机游走矩阵. 其中, 组合拉普拉斯矩阵的特征向量光滑性最好, 可用于

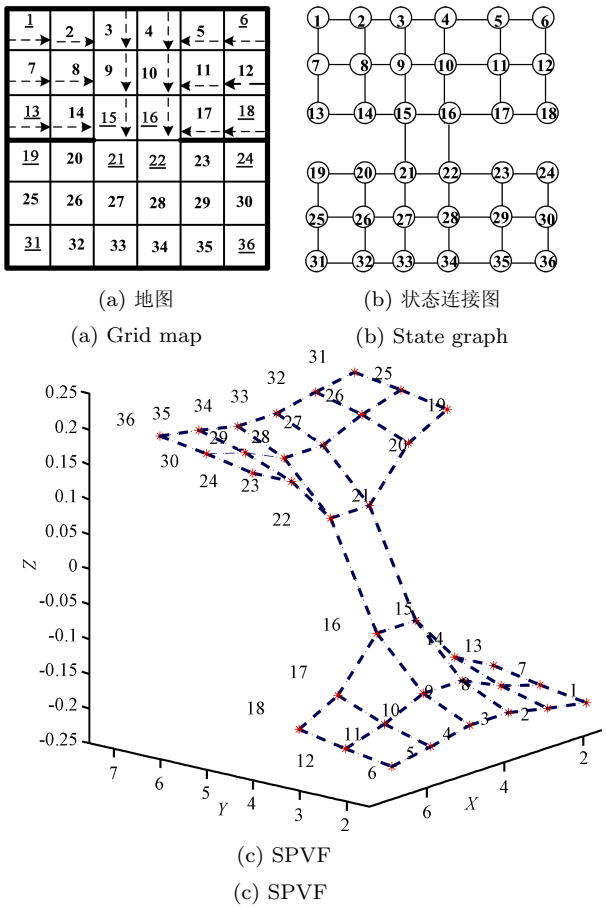
求取到达瓶颈状态的相关策略, 所以此处采用它来阐述 PVF 方法的基本思想.

Agent 通过随机游走或根据任务的背景知识可以建立一个表示状态空间邻接关系的无向图: $G = (V, E, W)$. 其中, V 为图顶点集合, E 为边集合, W 为边上的权值集合. 例如, 在图 2 (a) 所示的两个房间的 6 × 6 格子世界中, 中间黑粗线表示墙, 其状态连接图如图 2 (b) 所示, 状态相邻的边权值均为 1. 图 G 的组合拉普拉斯算子定义如下:

$$L = D - W \tag{10}$$

其中, D 为对角的度矩阵, $D(s, s') = \sum_{s \sim s'} w_{ss'}$, s 和 $s' \in V$, $s \sim s'$ 表示相邻顶点. 设函数 f 为图上每个顶点到实值的映射, $f : V(G) \to \mathbf{R}$. 组合拉普拉斯算子作用于函数 f 时定义为

$$Lf(s) = \sum_{s \sim s'} w_{ss'}(f(s) - f(s')) \tag{11}$$



(a) 地图 (b) 状态连接图

(a) Grid map (b) State graph

(c) SPVF

(c) SPVF

图 2 两个房间的 6 × 6 格子世界的图描述

Fig. 2 Graph description of the 6 × 6 grid with two rooms

求解方程 $Lf = 0$ 得到的函数即为调和函数, 文献 [17] 已证明求解调和函数等同于求解方程 $Lf = \lambda f$ 的特征向量 f. f 是一个特殊的图函数, 称为原型值函数[8], 其对应的特征值 λ 越小, 表明该函数越光滑. 非零最小特征值 λ_2 对应的 f_2 最光滑, 称为 Fiedler 特征向量或 SPVF, 该向量常用于图分割. 通过求取图拉普拉斯算子的特征向量可以形成一组正交的基函数, 用于逼近图上任意平方可积的函数, 包括强化学习中的值函数. 上述求解拉普拉斯矩阵的特征向量的方法在流形学习、谱聚类里被称为拉普拉斯特征映射法或谱方法[12−13].

在傅里叶分析中, 一个函数可以用一系列不同频率的正余弦函数之和来逼近, 并且低频部分代表了这个函数的主要特征. 类似于傅里叶分析, 在利用 PVF 进行值函数逼近时, 图拉普拉斯矩阵的小特征值形成的基函数的作用大于大特征值形成的基函数. 通常, 只需使用多个小特征值形成的基函数即可实现特定精度下的值函数逼近, 达到降维目的[8]. 此处把小特征值形成的基函数称为低频基函数.

根据谱图理论, SPVF 能自动捕获图的瓶颈和对称性等拓扑结构, 所以 PVF 方法可用于状态空间的层次分解[9−10]. 图 2 (c) 给出了两个房间的 6 × 6 格子世界的 SPVF, 图中的数字为状态编号, 虚线表示两个状态相邻, Z 轴为 SPVF 值. 从图可知, SPVF 值的分布是上下对称的 (以零为轴), 清楚地反映出 "门" 的位置和房间的对称性 (参见图 2 (a)). 通过寻找 SPVF 中相邻状态间差值最大的边就可找到瓶颈状态 15、16、21 和 22, 切断 15 与 21、16 与 22 之间的边就能实现任务的分解.

2 基于 PVF 的任务分解

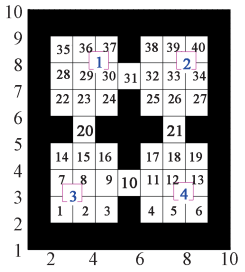
近年来, 把复杂问题分解成若干个小问题的分层强化学习取得显著进展, 先后提出了 Option、HAM 和 MAXQ 三种典型的方法[4]. 其中, Option 方法使用时态抽象技术, 将完成子任务的状态动作序列抽象成一个 Option, 把单步动作拓展到多步的情况形成宏动作, 减少了决策次数[18]. 如图 2 (a) 所示的 6 × 6 格子世界中, 门两边的状态 $\{15, 16, 21, 22\}$ 为瓶颈状态. 可以定义两个到门的子任务 Option, 在地图的上半部分定义 Option 1, 当 Agent 处于 1 ∼ 18 中的某个状态时, 执行图中箭头所示的策略, 直到到达终止状态 15 或者 16. 地图下半部分的到门 Option 2 的情况与此类似.

在早期研究中, 子任务对应的 Option 都是事先定义的. 近年来, 自动寻找子任务 Option 成为研究热点, 常用方法有状态访问频率法、值函数梯度法和谱图分割法等[9]. 谱图分割法巧妙地将任务的分解问题转化为图分割问题, 通过求取图的 PVF 来实现子图划分. 图分割时, 分割点所对应的状态就是瓶颈状态, 分割后得到的子图即为子任务.

2.1　基于谱图分割法的分层强化学习

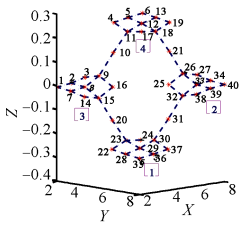
谱图分割法有坚实的理论基础, 对于稀疏图的分割效率较高, 适用于强化学习状态空间的分解. 文献 [10] 首先将此方法引入分层强化学习中, 利用规范割 (Normalized cut) 准则提出了 L-cut 算法. L-cut 算法是一种二分方法, 可方便地将任务分解为两个子任务. 但将其应用于复杂任务时, 需要迭代运行, 手工设置的参数较多, 计算复杂. 针对此缺点, Chung 等提出了一种级联分解算法[9]. 该算法直接利用任务的 SPVF 来确定多个瓶颈状态, 可一次将任务分割为多个子任务. 级联分解算法的缺点是需要事先设定子任务的分割数目, 本节后面将引入策略死锁检测机制来解决这个问题.

下面以四个房间的格子世界为例来说明级联分解算法的工作过程. 图 3 (a) 为所用例子的地图, 其中每个房间是一个子任务, 图 3 (b) 为相应的 SPVF. 级联分解算法直接利用式 (12) 寻找 SPVF 中相邻状态梯度绝对值的最大量来确定瓶颈状态 s_{key}, 迭代四次后就可以找到四个瓶颈状态: 10, 20, 21 和 31.



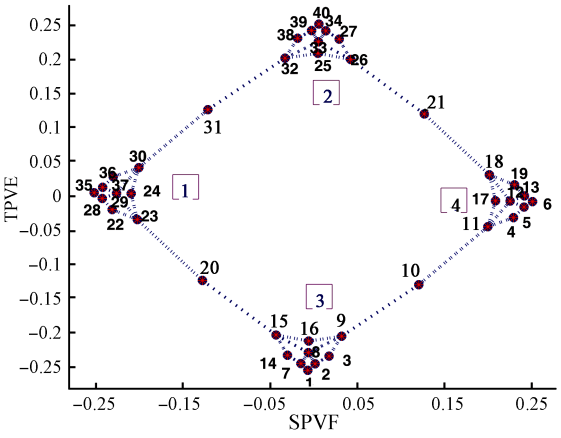
(a) 地图

(a) Grid map



(b) SPVF

(b) SPVF



(c) 状态空间的二维 PVF 表示

(c) Representation of state space with two PVFs

图 3　四个房间的 6×6 格子世界的图描述

Fig. 3　Graph description of the 6×6 grid with four rooms

切断与瓶颈状态相连的边就能完成图的分割, 获得四个子任务.

$$\begin{cases} \text{dif}f(s) = \max\limits_{s\sim s'}(\text{abs}(f_2(s) - f_2(s')) \\ s_{\text{key}} = \max\limits_{s\in G}(\text{dif}f(s)) \end{cases} \tag{12}$$

2.2　虚拟值函数法

文献 [9−10] 均采用了 SPVF 来寻找瓶颈状态以实现任务分解, 但子任务中到达瓶颈状态的相关策略还需重新学习. 事实上, 低频 PVF 可看作是状态空间至实数空间的一种映射, 映射后的像 (即 PVF 的元素) 保持了状态空间的局部拓扑, 各像元素间的欧氏距离可用于求取子任务的策略. 在简单层次任务中, 直接使用 SPVF 就可得到子任务的相关策略; 对于更复杂的情况, 需要先利用 n 个最光滑的 PVF 实现状态空间至 n 维实数空间的映射. 然后, 再计算相应像间的欧氏距离来确定策略.

$$\psi(s) = \sqrt{\sum_{i=2}^{n+1}(f_i(s_{\text{key}}) - f_i(s))^2} \tag{13}$$

$$\begin{cases} s_{\min} = \arg\min\limits_{s\sim s', s\neq s'}(\psi(s')) \\ \pi(s) = \arg\max\limits_{a\in A} p(s, a, s_{\min}) \end{cases} \tag{14}$$

利用 PVF 求取子任务中到达瓶颈状态 s_{key} 的策略时, 首先要根据任务的复杂程度选取所用低频 PVF 的数目 n; 然后, 利用式 (13) 计算子任务中各状态相对于瓶颈状态的虚拟状态值函数 $\psi(s)$; 最后, Agent 在状态 s 中使用一步搜索算法找到具有最小虚拟状态值函数的相邻状态 $s_{\min}$, 将 s 到 $s_{\min}$ 的最大概率动作作为 s 的子任务策略 (参见式 (14)). 本文把上述确定子任务策略的方法称为虚拟值函数法.

在虚拟值函数法中, 需要根据经验来选择所用低频 PVF 的数目 n. 如果 n 选得较小, 可能无法得到正确的子任务策略; n 较大, 则会增加计算量. 通常, 当任务的瓶颈状态间不存在环路时, 选择 $n = 1$ (即只用 SPVF); 对于存在环路的情况, 选择 $n = 2$. 例如, 在两个房间的格子世界中, 直接利用 SPVF 使用虚拟值函数法就可确定到达瓶颈状态的相关策略 (参见图 2 (c)). 但对于四个房间的格子世界问题, 则先需要利用 SPVF 和 TPVF (The third PVF) 将任务的状态空间映射到二维实数空间, 然后再求取各房间的子任务策略 (见图 3 (c)).

虚拟值函数法可看作是一类降维的人工势场法, n 通常都小于状态空间的维数. 人工势场法的一个缺点是易陷入局部极值点. 利用虚拟值函数求取相关策略时, 若子任务划分不当也会出现类似的情况 — 策略死锁 (也被称为极限环). 当检测到策略死锁

时, 可利用式 (12) 对相应子任务进行再划分, 直到消除死锁为止. 例如, 在四个房间的格子世界中, 若其只被划分为两个子任务: 房间 4 为子任务 1, 房间 1 ∼ 3 连接起来为子任务 2 (在图 3 (c) 中相当于切断 18 与 21, 9 和 10 之间的边). 在子任务 2 中, 当用虚拟值函数法求取到达瓶颈状态 21 的策略时, Agent 会在房间 3 中的状态 16 和 9 之间 "打转", 形成策略死锁 (状态 16 的策略是向下到达状态 9, 状态 9 的策略是向上到达状态 16). 这时就需对子任务 2 进行再划分. 在虚拟值函数法中, 使用策略死锁检测机制可以自适应地决定子任务的划分数目, 弥补了级联分解算法的不足.

2.3 改进的 PVF 层次任务分解方法

通过上述分析, 将基于 SPVF 的级联分解法、虚拟值函数法和策略死锁检测机制结合起来, 可得到一种改进的 PVF 层次任务分解方法. 其具体流程如下:

步骤 1. 根据强化学习任务的状态邻接关系, 建立环境的状态图论描述 $G = (V, E, W)$. 若状态 s 与 s' 相邻, 则其权值 $w_{ss'} = w_{s's} = 1$, 其余为零.

步骤 2. 由式 (10) 求得组合拉普拉斯算子, 然后计算相应的特征向量.

步骤 3. 重复利用式 (12) 寻找瓶颈状态 s_{key}, 切断 s_{key} 之间的边来分割图 G 以得到相应的子任务. 划分子任务数目的初始值由领域知识决定.

步骤 4. 根据任务的复杂度来选择所用低频 PVF 的数目 n, 然后采用式 (13) 和式 (14) 所示的虚拟状态值函数法来确定子任务的相关策略, 形成子任务 Option.

步骤 5. 检查各子任务 Option 的策略是否存在死锁现象. 若存在, 则利用 SPVF 对该子任务的图进行再划分, 直至消除死锁为止.

改进的层次任务分解方法合理地利用了拉普拉斯特征映射能保持状态空间局部拓扑结构不变的特点, 增加了策略死锁检测机制来避免可能出现的策略错误, 可自适应地确定子任务的划分数目. 需要说明的是, 方法中若使用正则化拉普拉斯矩阵或随机游走矩阵时, 不能保证得到的子任务策略是最优的. 原因是它们产生的 SPVF 的光滑性相对较差, 不能反映状态空间的局部拓扑结构.

3 基于 PVF 的迁移

正如第 1.3 节所述, PVF 方法需要访问每个状态以得到状态空间的连接图, 对于状态空间较大的问题采样成本较高, 所以引入迁移学习. PVF 方法产生的基函数独立于回报函数, 对于状态连接图变化很小的任务, 源任务产生的基函数可直接用于目标任务的学习中. 对于地图比例缩放的任务, 则需处理基函数维数不一致的问题, 文献 [11] 给出的方法是 Nyström 扩展和线性插值技术. Nyström 扩展仍然需要目标任务状态空间的连接信息, 只是减小了特征值的计算维数, 并未降低采样复杂度. 线性插值方法在状态空间比例缩小的任务中效果很好, 但在比例放大因子较大的任务中效果不理想, 文献 [11] 对此并未做深入讨论. 本节将以两个房间的格子世界为例, 从插值所得基函数的逼近能力和正交性角度出发, 讨论线性插值方法在比例放大任务中不稳定的原因, 并分析插值所得的近似 PVF 用于目标任务层次分解的可能性. 在此基础上, 提出基函数与层次结构相结合的混合迁移方法.

3.1 基函数的线性插值

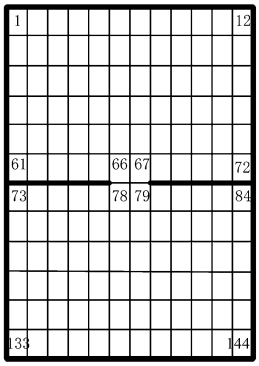
线性插值技术主要用于解决基函数维数不一致的问题, 下面以格子世界为例阐明其原理. 首先, Agent 通过沿墙行走找到迁移任务的角状态和临墙状态. 然后, 将源任务和目标任务的角状态一一对应. 最后, 根据任务临墙状态的比例关系, 对源任务的特征向量进行线性插值或删值, 以得到目标任务的基函数[11]. 在 6×6 地图到 12×12 地图的迁移任务中, 将图 2 (a) 中带下划线并斜体标示的角状态与图 4 (a) 中的角状态一一对应, 目标任务增加的行列状态的基函数值由源任务的基函数线性插值得到, 图 2 (a) 中瓶颈状态所在行和列间不插值.

PVF 方法的抽象傅里叶特性决定了相似任务间只有少数低频基函数是相似的, 线性插值技术并未改变这个性质. 所以, 扩展得到的基函数与目标任务的精确基函数只在低频部分形状相似, 在高频部分有较大差别. 图 4 (b) 对比了 12×12 地图的精确 SPVF 与由 6×6 地图插值形成的近似 SPVF, 为了可视化需要, 将插值 SPVF 的所有元素都乘以 -1. 从图中可知: 两类 SPVF 几乎一样, 保持了相同的结构, 都能用于目标任务的层次分解和确定子任务的最优 Option. 图 5 给出了第 3 ∼ 6 个低频基函数的对比情况, 从中可以看出低频基函数非常相似. 从第 15 个基函数开始, 两类 PVF 出现较大差别, 由于篇幅有限这里不再列图给出.

PVF 方法产生的基函数都是线性无关且正交的. 使用线性插值后, 所得基函数的线性无关性可以保持, 但高频部分的正交性会被破坏. 对由 6×6 地图插值得到的 12×12 地图的基函数进行测试发现, 只有前 10 个基函数保持相互正交.

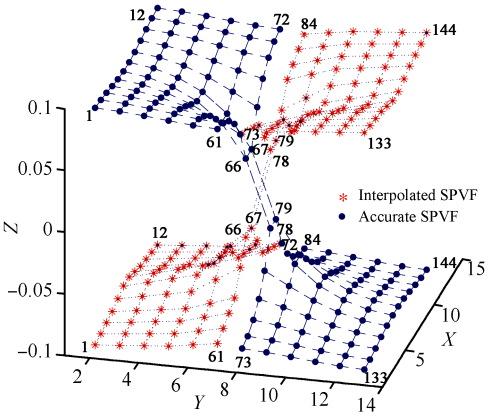
3.2 迁移基函数的逼近能力

强化学习最终的目的是获取最优策略, 求取值函数只是中间步骤. 根据广义策略迭代思想[1], 策略评估中不必求取当前策略的精确值函数, 只要所求值函数能保证贪婪改进后得到更优的新策略即可. 对



(a) 地图

(a) Grid map



(b) SPVF 对比

(b) Comparison of two SPVFs

图 4　12 × 12 格子世界地图及其 SPVF

Fig. 4　A 12 × 12 grid and its SPVF

于同一个强化学习问题, 若数值不同的值函数贪婪改进后得到策略是相同的, 就认为两个值函数包含了同样的策略结构, 定义为策略一致.

根据 PVF 方法能保持全局光滑性的特点, 在同一问题中使用不同数目的 PVF 构成的逼近器求得的值函数是不同的, 但只要其与精确值函数策略一致, 就认为该逼近器有效. 通常, 用较少的低频基函数逼近的值函数能保持精确值函数的全局形状, 但局部状态存在逼近错误, 不能保证策略一致性. 例如在 6 × 6 任务中, 若目标状态为图 2 (a) 中的状态 1, 其最优值函数如图 6 的左栏所示. 图 6 的中栏为用最光滑的 6 个低频基函数构成线性逼近器, 在模型和最优策略已知情况下使用不动点法 (参见式 (8)) 求得的值函数. 从左栏与中栏图的对比可知: 后者的值函数的多个角状态值不准确, 并未体现出给定的最优策略. 基函数不足的逼近器在模型已知情况下尚且有逼近错误, 用于策略迭代时更会将此类错误累计放大, 最终导致算法无法求得最优策略.

在状态空间比例放大的 PVF 迁移任务中, 插值得到的基函数数目会比目标任务的精确基函数数目少, 并且基函数间不能保证相互正交, 用于逼近目标任务的值函数时会出现严重的逼近错误. 例如, 在 6 × 6 地图迁移到 12 × 12 地图的任务中, 目标任务需要 144 个基函数才能保证准确地逼近状态空间中任何平方可积的值函数. 而由 6 × 6 任务插值只能得到 36 个基函数, 保持正交的只有 10 个, 其逼近能力非常有限. 图 6 的右栏图为 12 × 12 地图中用最光滑的 10 个插值基函数在最优策略下逼近的值函数的情况, 图中有较多角状态的值函数不正确 (12 × 12 地图的最优值函数与左栏图类似). 比例放大因子越大, 此类情况越严重, 这就是为什么线性插值方法在状态空间比例缩小任务中效果很好, 而在比例放大任务中不稳定的原因.

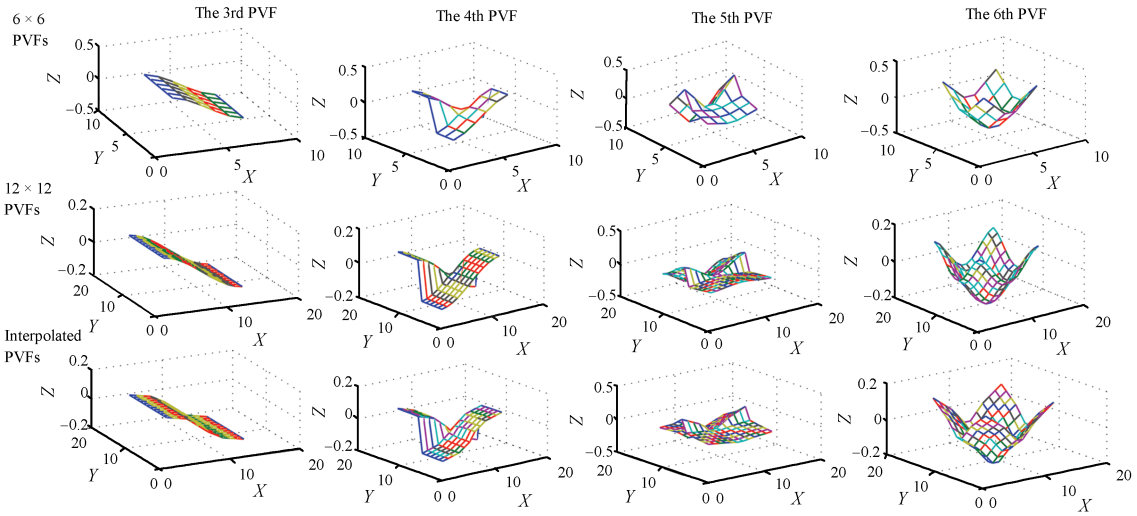


图 5　第 3 ∼ 6 个 PVF 对比情况

Fig. 5　Comparision of the 3rd ∼ 6th PVFs

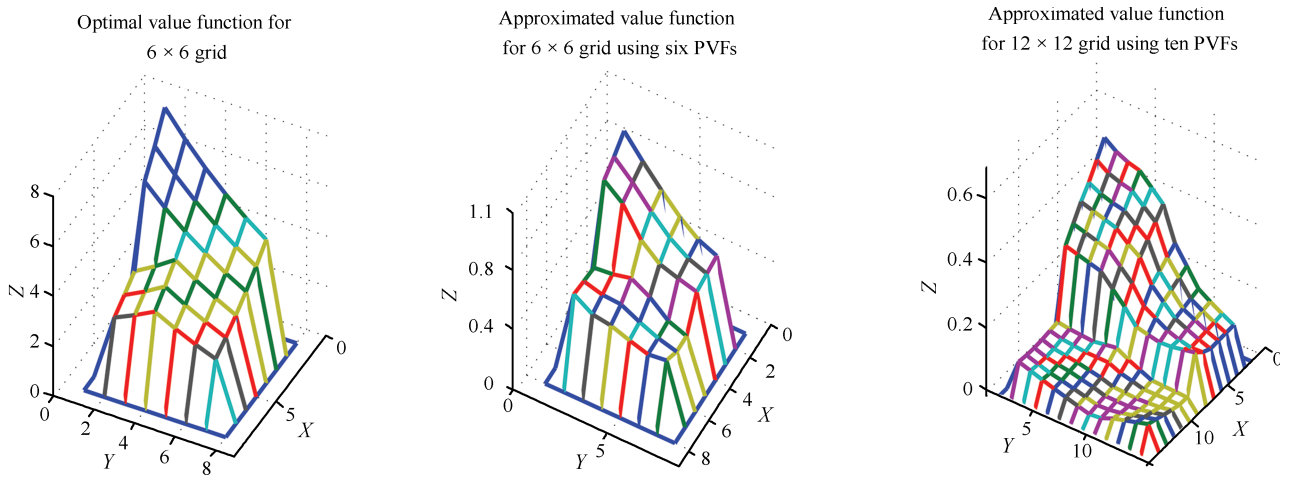


图 6 值函数比较

Fig. 6 Comparison of value function

3.3 基函数与子任务策略的混合迁移

强化学习中由基函数数目较少产生的逼近错误会在策略迭代过程中累计放大, 不仅增加了算法的迭代次数, 还会使最终策略不满足最优. 简单有效的解决方法是增加基函数数目, 这在地图比例放大的迁移任务中显然不可行.

PVF 方法中高频基函数决定被逼近函数的局部细节, 由数量较少的低频基函数产生的逼近错误通常出现在值函数梯度较小的部分, 特别是离目标状态较远的角状态. 修正这部分状态的错误策略就可获得最优策略. 在 PVF 迁移中, 线性插值得到的近似 SPVF 与精确的 SPVF 保持同样的结构, 也可用于任务分解. 因此, 可以把对齐的低频基函数和近似 SPVF 得到的子任务 Option 结合起来迁移, 利用第 2.3 节所述的改进层次分解方法, 直接确定目标任务中部分状态空间的最优策略. 直接确定最优策略的状态区域通常离目标位置较远, 与值函数逼近错误的区域较为一致 (参见图 6 (c)), 所以改进的迁移方法可以修正大部分由逼近错误导致的错误策略.

4 混合迁移方法步骤

基于以上分析, 针对状态空间具有层次结构的强化学习任务, 本文所提出的混合迁移方法的主要步骤如下:

步骤 1. 首先, 根据源任务的状态邻接关系, 建立起环境的状态图论描述 G, 并求得组合拉普拉斯算子 L_S. 然后, 通过计算 L_S 的特征值向量来产生源任务的基函数 Φ_S. 最后, 依据式 (12) 利用 f_{S2} (源任务的 SPVF) 来寻找 s_{key}.

步骤 2. 在目标任务 D_T 中, 首先根据采样得到的状态–动作序列和任务的领域知识, 确定目标任务的关键状态和地图间的比例关系. 然后, 使用线性插值技术将基函数 Φ_S 扩展成 Φ_T.

步骤 3. 取出扩展 Φ_T 中的近似 f_{T2}, 利用第 2.3 节所示的 PVF 层次任务分解方法中的步骤 3 ∼ 步骤 5 来求取相关子任务的 Option.

步骤 4. 根据目标任务的背景知识和目标状态所在的位置, 从上一步骤得到的子任务 Option 中选取合适的 O_{tran} 用于迁移.

步骤 5. 将对齐的基函数 Φ_T 和选取的 O_{tran} 一起用于目标任务的学习算法中 (LSPI 和 Q 算法都可以). 在策略迭代时, 包含在 O_{tran} 中的状态一直使用自身的子任务最优策略, 其他状态采用设定的策略改进方式来更新策略.

上述方法假设两个任务状态空间的拓扑是完全一致的, 并且易于对齐 (很多时候采用广义对齐技术), 这点在实际的任务中较难满足. 在实际应用中, 可以在目标任务的学习过程中逐步完善状态连接图, 并验证其与源任务连接图的相似性. 同时, 在步骤 5 中采用类似于 ε-greedy 的方式选择子任务策略, 以弥补由于任务间的拓扑不完全一致或存在对齐误差而引起的错误.

5 仿真结果及分析

为了验证所提迁移方法的有效性, 本文分别在两个房间和四个房间的格子世界中进行了仿真研究. 仿真中, Agent 有向东、向西、向南、向北四个确定性动作, 碰墙后 Agent 状态不变, 非目标状态的立即回报值函数为 0, 目标状态的为 1.

为了提高效率, 减小采样样本数量和迭代顺序对算法比较的影响, 仿真中采用式 (8) 所示的模型已知的 LSPI 算法, 所用的基函数由状态空间的基函数直接复制到动作空间产生[14]. 仿真中对比了基

函数迁移 (Basis function transfer, BFT) 和混合迁移 (Hybrid transfer, HT) 的效果. 使用基函数迁移时, 源任务与目标任务之间仅迁移基函数, 所有状态都采用贪婪方式改进策略; 采用混合迁移时, 任务之间同时迁移基函数和子任务 Option, 属于子任务 Option 的状态在策略迭代中使用 Option 的策略, 剩余的状态采用贪婪方式改进策略.

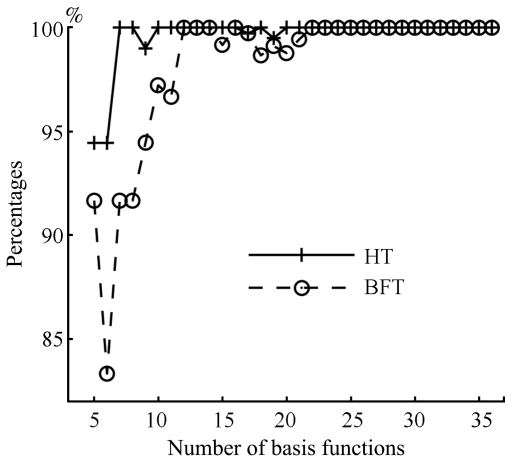
所有仿真中, 目标任务的目标状态位置在地图的左上角, 学习算法中 $\gamma = 0.88$. 策略迭代中, 当 Q 值变化量的绝对值都小于 0.0001 时, 认为策略收敛. 学习结束时, 若某状态求得的策略与由任务的精确值函数确定的策略一致, 则认为算法在该状态下取得了最优策略. 在仿真的例子中, 有些状态由精确值函数确定的策略可以是多个动作, 若算法求得策略与其中一个动作相同, 就认为得到了最优策略. 每组实验随机运行 50 次, 结果中的最优策略成功率指算法取得最优策略的状态数目与状态空间的总状态数 (即 $|S|$) 的比率, 将多次实验的最优策略成功率取均值就得到最优策略成功率的平均值.

5.1　地图不变迁移

在地图不变的迁移仿真中, 源任务和目标任务的基函数完全一致, 无需使用线性插值技术. 实验分别使用了两个房间的 6×6、8×8、12×12 的地图 (参见图 2 和图 4) 和四个房间的 6×6 的地图 (每个房间大小为 3×3, 见图 3)、10×10 地图 (每个房间大小为 5×5). 每次运行时, 两种迁移方法的初始随机策略相同.

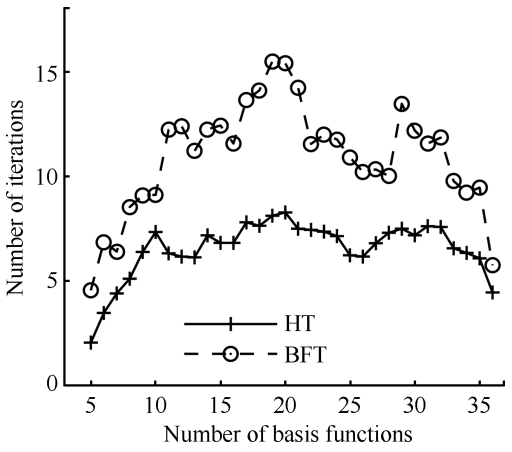
两个房间的 6×6 地图仿真中, 基函数的数目变化范围为 5～36, 策略迭代的最大代数为 30. 图 7 (a) 对比了两种迁移方法的最优策略成功率的平均值, 结果表明: HT 可有效提高最优策略成功率, 减少值函数逼近所必需的最小基函数数目 (在图中对应于最优策略成功率的平均值为 100 %). 图 7 (b) 为平均策略迭代次数对比图, 从中可以看出, HT 策略迭代次数明显少于 BFT 的情况. 其原因是 HT 中大约一半状态空间的最优策略已确定, 降低了策略迭代的开销. 两个房间的 8×8、12×12 地图的仿真结果与 6×6 地图类似, 这里不再赘述.

在图 3 所示的四房间格子世界的仿真中, 任务的目标位置在房间 1. 考虑到在房间 4 中, 由虚拟值函数法求取的到门策略不一定是任务的最优策略 (房间 4 不与目标所在的房间 1 相邻, 房间 4 的最优策略与目标在房间 1 的位置有关), 我们分别做了两个房间和三个房间的策略迁移仿真, 用于对比不同情况下所提方法的迁移效果. 在两个房间的策略迁移仿真中, 房间 2 和房间 3 的策略使用由虚拟值函数法求得的到门策略, 并在策略迭代中固定不变, 房间 1 和房间 4 的策略由学习得到. 三个房间的策略迁移实验中, 房间 2～4 的策略固定, 只需学习房间 1 的策略. 房间 2 和房间 3 使用的策略仍是到门策略, 房间 4 的策略设定为任务的最优策略. 两种策略迁移的区别是三房间策略迁移可以修复值函数梯度最小的房间 4 的策略错误, 而二房间策略不能修复.



(a) 最优策略成功率的平均值

(a) Average of optimal policies′ success rate

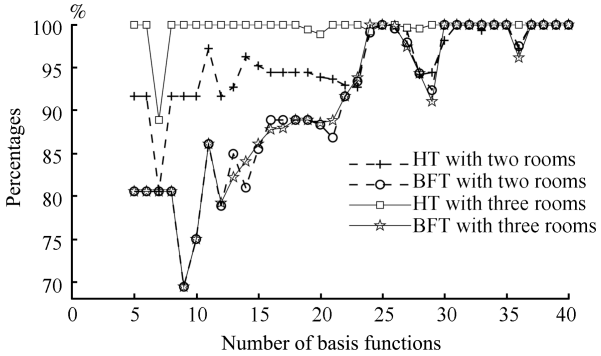


(b) 收敛平均迭代次数

(b) Average iterations for convergence

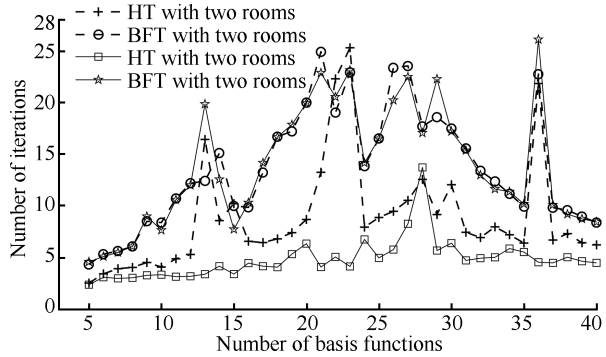
图 7　两个房间的 6×6 地图仿真结果

Fig. 7　Simulation results of 6×6 grid with two rooms



(a) 最优策略成功率的平均值

(a) Average of optimal policies′ success rate



(b) 收敛平均迭代次数

(b) Average iterations for convergence

图 8 四个房间的 6 × 6 地图仿真结果

Fig. 8 Simulation results of 6 × 6 grid with four rooms

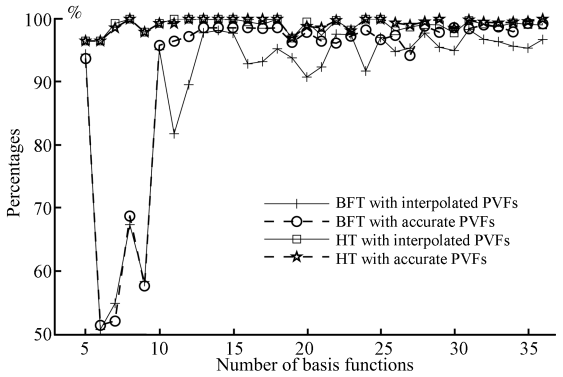
图 8 为四个房间的 6 × 6 地图的仿真结果, 从中可以看出: 1) 在较复杂的例子中, 论文所提的混合迁移方法仍然有效; 2) 两房间策略的混合迁移方法相对于二房间和三房间的基函数迁移来说, 并未明显减少值函数逼近所必需的最小基函数数目, 但学习效率仍然有所提高; 3) 三房间策略迁移结果的学习效率和稳定性均好于两房间的. 四个房间的 12 × 12 地图的仿真结果与之类似.

5.2 地图比例放大迁移

本节的仿真中, 目标任务和源任务地图的拓扑结构相似, 但状态空间大小不一样. 两个房间的例子中, 源任务的地图大小 6 × 6, 目标任务为 8 × 8 和 12 × 12. 四个房间的例子中, 源任务为 6 × 6 的地图, 目标任务为 10 × 10 的地图. 文中所述插值基函数是用线性插值方法扩展源任务的基函数而形成的, 精确基函数是指由目标任务的状态连接图计算得的基函数, 用于与插值基函数仿真结果做比较.

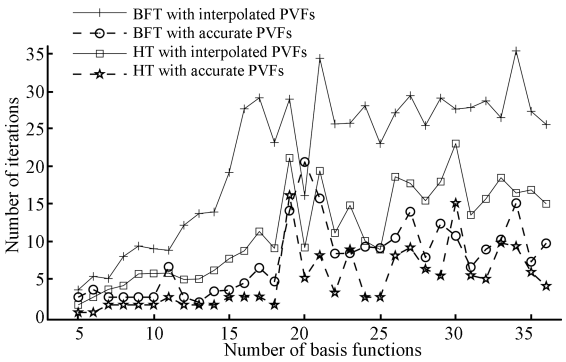
图 9 是目标任务为两个房间的 12 × 12 地图的仿真结果. 图 9 (a) 对比了两种迁移方法的最优策略成功率的平均值, 从中可知: 1) 纯粹迁移基函数的方法由于函数逼近能力的限制使得最优策略成功率的平均值较低, 而采用混合迁移时有明显提高. 这也说明混合迁移时, 子任务 Option 的策略成功地修正了所在子区域的错误策略. 2) 插值基函数与精确基函数在基函数数目较小时, 最优策略成功率较为一致, 但在数目较大时有差别. 其主要原因是插值基函数在高频部分形状与精确基函数有差别, 且不再保持正交. 图 9 (b) 为不同迁移方法的平均策略迭代次数比较图, 结果与预期一样: 精确基函数的平均迭代次数小于插值基函数的, 混合迁移的情况小于纯粹基函数迁移的情况.

图 10 对比了目标任务为四个房间的 10 × 10 地图的精确二维 PVF 表示和插值二维 PVF 表示, 从中可以看出两者有相似的拓扑关系, 所以插值 PVF 也能用于子任务策略的求取. 四个房间的 10 × 10 地图的仿真结果与前述的 12 × 12 地图例子类似, 结果参见表 1 ∼ 3.



(a) 最优策略成功率的平均值

(a) Average of optimal policies′ success rate



(b) 收敛平均迭代次数

(b) Average iterations for convergence

图 9 两个房间的 12 × 12 地图仿真结果

Fig. 9 Simulation results of 12 × 12 grid with two rooms

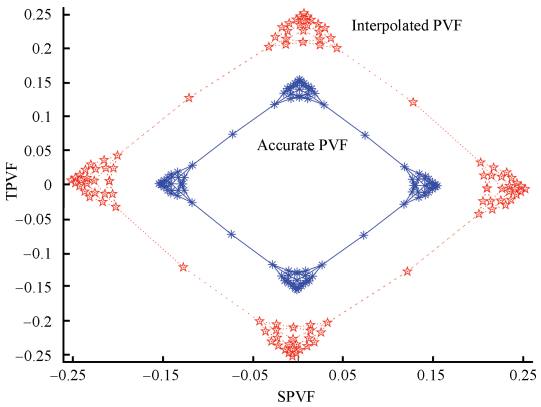


图 10 状态空间的二维 PVF 表示的对比图

Fig. 10 Representation of state space with different PVFs

为了更直观地说明混合迁移方法的仿真效果, 表 1 ∼ 3 对前面所述的各类实验进行了统计 (表中的 “∗” 表示三房间策略迁移). 其中, 表 1 统计了各类

实验中求取到全部状态的最优策略时所用的最少基函数数目, 表 2 统计了不同的基函数数目对应的收敛平均迭代次数的均值, 表 3 统计了不同的基函数数目对应的平均最优策略成功率的均值. 表 1～3 的结果对比清楚地表明: 混合迁移方法可以有效减少求取全部状态最优策略所需的最少基函数数目, 降低策略迭代次数, 提高平均最优策略成功率, 较好地弥补了状态空间比例放大任务中的有效基函数不足的缺点.

6 结论

在基于 PVF 方法的基函数迁移研究中, 常用的线性插值技术可以将小状态空间的源任务基函数扩展为大状态空间的目标任务的基函数, 但扩展后的基函数用于目标任务时会出现逼近错误. 本文首先分析了线性插值所得基函数的特点和正交性, 讨论了近似 PVF 用于任务分解的可能性; 然后, 基于改进的谱图理论层次分解技术, 提出了一种基函数与子任务 Option 相结合的混合迁移方法. 对于状态空间具有层次结构的任务, 改进的混合迁移方法可以直接确定目标任务部分状态空间的最优策略. 最后, 在格子世界中对提出的方法进行了仿真研究. 仿真结果表明, 该方法可有效减少求取最优策略所需的最少基函数数目, 降低策略迭代次数. 所提的混合迁移方法不能解决与目标状态在同一子区域的值函数逼近错误问题, 改进的层次分解法在包含多门和不规则墙的复杂地图应用中也无法保证得到全局最优策略. 下步工作将重点解决以上不足并将其扩展到具有连续状态空间与动作空间的任务中. 另外, 本文的工作假设任务的状态图是易于对齐的, 这点在实际任务中不易满足, 后续工作会对此深入研究.

表 1 求取最优策略所需最小基函数数目对比

Table 1 Minimum numbers of basis functions needed for optimal policies

地图	两房间格子世界				四房间格子世界			
源任务	6×6	12×12	6×6	6×6	6×6	$6\times6^*$	6×6	$6\times6^*$
目的任务	6×6	12×12	8×8	12×12	6×6	$6\times6^*$	10×10	$10\times10^*$
精确基函数	12	37	16	37	30	30	>40	>40
精确基函数 (混合)	6	12	7	12	30	8	>40	22
插值基函数	−	−	13	>36	−	−	>40	>40
精确基函数 (混合)	−	−	9	11	−	−	>40	24

表 2 总平均迭代次数对比

Table 2 Total average iterations

地图	两房间格子世界				四房间格子世界			
源任务	6×6	12×12	6×6	6×6	6×6	$6\times6^*$	6×6	$6\times6^*$
目的任务	6×6	12×12	8×8	12×12	6×6	$6\times6^*$	10×10	$10\times10^*$
精确基函数	11.1	31.7	4.1	6.9	13.8	13.9	15.5	13.2
精确基函数 (混合)	6.7	18.7	2.9	4.5	8.9	4.8	17.8	4.4
插值基函数	−	−	14.3	22.0	−	−	36.4	33.1
精确基函数 (混合)	−	−	8.2	11.9	−	−	26.7	15.8

表 3 最优策略成功率的平均值对比

Table 3 Average success rates of optimal policies

地图	两房间格子世界				四房间格子世界			
源任务	6×6	12×12	6×6	6×6	6×6	$6\times6^*$	6×6	$6\times6^*$
目的任务	6×6	12×12	8×8	12×12	6×6	$6\times6^*$	10×10	$10\times10^*$
精确基函数	98 %	92 %	93 %	93 %	91 %	89 %	90 %	>40
精确基函数 (混合)	99.6 %	99.2 %	99.9 %	99 %	96 %	99 %	94 %	98 %
插值基函数	−	−	90 %	90 %	−	−	88 %	89 %
精确基函数 (混合)	−	−	99.6 %	99 %	−	−	93 %	98 %

References

1 Sutton R S, Barto A G. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998

2 Gao Yang, Chen Shi-Fu, Lu Xin. Research on reinforcement learning technology: a review. *Acta Automatica Sinica*, 2004, **30**(1): 86−100
(高阳, 陈世富, 陆鑫. 强化学习研究综述. 自动化学报, 2004, **30**(1): 86−100)

3 Zhao Dong-Bin, Liu De-Rong, Yi Jian-Qiang. An overview on the adaptive dynamic programming based urban city traffic signal optimal control. *Acta Automatica Sinica*, 2009, **35**(6): 676−681
(赵冬斌, 刘德荣, 易建强. 基于自适应动态规划的城市交通信号优化控制方法综述. 自动化学报, 2009, **35**(6): 676−681)

4 Barto A G, Mahadevan S. Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, 2003, **13**(4): 341−379

5 Pan S J, Yang Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010, **22**(10): 1345−1359

6 Taylor M E, Stone P. Transfer learning for reinforcement learning domains: a survey. *The Journal of Machine Learning Research*, 2009, **10**: 1633−1685

7 Wang Hao, Gao Yang, Cheng Xing-Guo. Transfer of reinforcement learning: the state of the art. *Acta Electronica Sinica*, 2008, **36**(12a): 39−43
(王皓, 高阳, 陈兴国. 强化学习中的迁移: 方法和进展. 电子学报, 2008, **36**(12a): 39−43)

8 Mahadevan S, Maggioni M. Proto-value functions: a Laplacian framework for learning representation and control in Markov decision processes. *The Journal of Machine Learning Research*, 2007, **8**: 2169−2231

9 Chiu C C, Soo V W. Automatic complexity reduction in reinforcement learning. *Computational Intelligence*, 2010, **26**(1): 1−25

10 Simsek O, Wolfe A P, Barto A G. Identifying useful subgoals in reinforcement learning by local graph partitioning. In: Proceedings of the 22nd International Conference on Machine Learning. New York, USA: ACM, 2005. 816−823

11 Ferguson K, Mahadevan S. Proto-transfer Learning in Markov Decision Processes Using Spectral Methods, Technical Report, University Massachusetts, Amherst, 2008

12 Luo Si-Wei, Zhao Lian-Wei. Manifold learning algorithms based on spectral graph theory. *Journal of Computer Research and Development*, 2006, **43**(7): 1174−1179
(罗四维, 赵连伟. 基于谱图理论的流形学习算法. 计算机研究与发展, 2006, **43**(7): 1174−1179)

13 Shi J B, Malik J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, **22**(8): 888−905

14 Lagoudakis M G, Parr R. Least-squares policy iteration. *Journal of Machine Learning Research*, 2003, **4**(12): 1107−1149

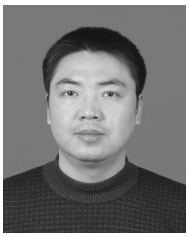
15 Wang Xue-Song, Tian Xi-Lan, Cheng Yu-Hu, Yi Jian-Qiang. Q-learning system based on cooperative least squares support vector machine. *Acta Automatica Sinica*, 2009, **35**(2): 214−219
(王雪松, 田西兰, 程玉虎, 易建强. 基于协同最小二乘支持向量机的 Q 学习. 自动化学报, 2009, **35**(2): 214−219)

16 Xu X, Hu D W, Lu X C. Kernel-based least squares policy iteration for reinforcement learning. *IEEE Transactions on Neural Networks*, 2007, **18**(4): 973−992

17 Chung F R K. *Spectral Graph Theory*. United States: American Mathematical Society, 1996

18 Sutton R S, Precup D, Singh S. Between mdps and semi-mdps: a framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 1999, **112**(1−2): 181−211

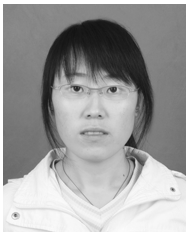
朱美强 中国矿业大学讲师, 博士研究生. 主要研究方向为机器学习, 智能优化与控制. 本文通信作者.
E-mail: zhumeiqiang@cumt.edu.cn
(**ZHU Mei-Qiang** Lecturer and Ph. D. candidate at China University of Mining and Technology. His research interest covers machine learning, intelligent optimization and control. Corresponding author of this paper.)



程玉虎 中国矿业大学教授. 主要研究方向为机器学习, 智能优化与控制.
E-mail: chengyuhu@163.com
(**CHENG Yu-Hu** Professor at China University of Mining and Technology. His research interest covers machine learning, intelligent optimization and control.)

李 明 中国矿业大学教授. 主要研究方向为智能控制, 模式识别.
E-mail: liming@cumt.edu.cn
(**LI Ming** Professor at China University of Mining and Technology. His research interest covers intelligent control and pattern recognition.)

王雪松 中国矿业大学教授. 主要研究方向为机器学习, 生物信息学.
E-mail: wangxuesongcumt@163.com
(**WANG Xue-Song** Professor at China University of Mining and Technology. Her research interest covers machine learning and bioinformatics.)



冯涣婷 中国矿业大学硕士研究生. 主要研究方向为强化学习.
E-mail: fhtcumt@163.com
(**FENG Huan-Ting** Master student at China University of Mining and Technology. Her main research interest is reinforcement learning.)