文章编号: 1001-0920 (2014) 03-0425-06 DOI: 10.13195/j.kzyjc.2012.1669

基于拉普拉斯特征映射的启发式Q学习

朱美强, 李 明, 程玉虎, 张 倩, 王雪松

(中国矿业大学 信息与电气工程学院，江苏 徐州 221116)

摘 要: 在基于目标的强化学习任务中, 欧氏距离常作为启发式函数用于策略选择, 其用于状态空间在欧氏空间内不连续的任务效果不理想. 针对此问题, 引入流形学习中计算复杂度较低的拉普拉斯特征映射法, 提出一种基于谱图理论的启发式策略选择方法. 所提出的方法适用于状态空间在某个内在维数易于估计的流形上连续, 且相邻状态间的连接关系为无向图的任务. 格子世界的仿真结果验证了所提出方法的有效性.

关键词: 强化学习；启发式策略选择；Q学习；拉普拉斯特征映射

中图分类号: TP181 **文献标志码:** A

Heuristically accelerated Q-learning algorithm based on Laplacian Eigenmap

ZHU Mei-qiang, *LI Ming*, *CHENG Yu-hu*, *ZHANG Qian*, *WANG Xue-song*

(School of Information and Electrical Engineering，China University of Mining and Technology，Xuzhou 221116, China．Correspondent：ZHU Mei-qiang，E-mail：tianlianglu@163.com)

Abstract: As a heuristic function, the Euclidean distance is usually used to select online action in reinforcement learning based on goal position. It is not applied to these tasks whose state spaces are not continuous in Euclidean space. For the problem, the Laplacian Eigenmap whose computational complexity is lower in manifold learning is introduced, then a method of heuristic policy selection based on the spectral graph theory is proposed. The proposed method is suitable for these tasks not only whose state spaces are continuous in some manifold that has a good estimation of intrinsic dimension, but also whose connection relation is expressed by an undirected graph. The simulation results of grid world show the effectiveness of the proposed method.

Key words: reinforcement learning；heuristic policy selection；Q-learning；Laplacian Eigenmap

0 引 言

强化学习 (RL) 能在无环境模型和教师样本的情况下, 通过与环境交互进行自主学习, 已广泛应用于调度优化、自适应控制和机器人自主导航等领域[1-5]. RL的主要缺点是学习效率较低, 原因在于其核心的试错改进机制和回报延迟的特点决定了智能体仅能依据学习中获取的稀疏回报来改进策略, 而忽略了大量有用的信息和知识. 从20世纪90年代起, 研究者开始抛弃智能体“一无所知”的假设, 通过发现和利用问题的领域知识提高RL的效率[1-2,4].

RL的算法较多, 主要包括Q学习、SARSA学习和R学习等, 其中Q学习应用最为广泛. 作为一种模型无关的在线时间差分 (TD) 学习方法, Q学习的策略选择方法直接影响算法的效率 (即探索和利用难题). 常用的策略选择方法有Boltzmann分布、ε-greedy、贝叶斯方法和启发式策略选择 (也称Action Biasing或Control Sharing) 等[1-4]. 这些方法中, Boltzmann分布和ε-greedy并未有效利用经验知识; 贝叶斯方法虽然理论坚实, 但存在采样和计算复杂、先验概率不易确定、未有效使用过程知识等缺点; 启发式策略选择更为灵活, 直接使用相关领域知识指导智能体的动作选择, 先验和过程知识都可以使用[6-8]. 文献[8]在小车爬山和倒立摆任务中对比研究了多类启发式强化学习方法, 结果表明启发式策略选择具有稳定的学习

收稿日期: 2012-11-06；**修回日期:** 2013-04-18.

基金项目: 国家自然科学基金项目(61072094, 61273143)；教育部高等学校博士学科点专项科研基金项目(20110095110011, 20110095110016)；中央高校基本科研业务费专项资金项目(2013XK09)；江苏省自然科学基金项目(BK20130207)；江苏省博士后基金项目(1301029C).

作者简介: 朱美强(1979−), 男, 讲师, 博士, 从事机器学习、智能控制的研究；李明(1962−), 男, 教授, 博士, 从事机器学习、机器人与智能控制等研究.

加速性能.

启发式策略选择中, 启发式函数的质量直接决定算法的性能, 通常为某类广义距离或者规则. 这些距离和规则既可以由问题的经验和领域知识直接确定, 也能从相似任务中迁移得到, 或者在算法运行过程中自学习产生[6-7]. 对于状态空间在欧氏空间内连续的任务, 欧氏距离常作为启发式函数, 对于迷宫这类状态空间在欧氏空间内不连续的任务效果不佳. 针对这类问题, 文献 [6] 提出了启发式加速强化学习算法 (HARL), 该算法通过在学习过程中利用结构提取和启发式信息反向传播技术得到建议策略. 首先探索整个状态空间, 得到估计的状态转移矩阵, 然后利用基本的动态规划算法实现启发式策略的反向传播, 计算较为复杂.

某种意义上讲, HARL 是一种基于模型的强化学习算法, 估计的传递矩阵包含了状态空间的邻接关系. 若将拉普拉斯特征映射法 (LE)[9]应用于由该邻接关系所建立的图中, 则不但可以有效提取状态空间的结构信息, 而且能实现状态空间降维和任务分解, 此思路已应用于值函数泛化和分层强化学习中[10-11]. 事实上, LE 也是一种计算效率较高的流行学习方法, 能在特定条件下将流形上连续的曲面在低维欧氏空间里“铺平”[5,9,11]. 对于状态空间在欧氏空间不连续、但在某个流形上连续的任务, 利用该方法能够在映射的低维欧氏空间里将状态空间“展开”, 用展开后的欧氏距离作为启发式函数即可避免直接使用欧氏距离的不足. 基于上述思想, 本文提出了一种基于 LE 的启发式策略选择方法, 并通过仿真实验表明了所提出方法的有效性.

1 Q 学习的启发式策略选择

Q 学习是应用最为广泛的 RL 算法, 其不用估计环境的模型, 直接利用下式所示的预测方法迭代求解动作值函数[1]:

$$Q(s,a)\rightarrow$$
$$Q(s,a)+\alpha[r(s,a,s')+\gamma\max_{a'\in A}Q(s',a')-Q(s,a)]. \quad (1)$$

其中: $\{a,a'\}\in A$, $\{s,s'\}\in S$, S 为有限状态空间, A 为有限动作空间, $\gamma\in(0,1]$ 为折扣因子, $\alpha\in(0,1]$ 为学习率, $r(s,a,s')$ 为在状态 s 执行动作 a 后转移到状态 s' 时得到的立即回报.

Q 学习是一类使用离线策略的在线 TD 学习方法, 更新 Q 值函数时用到的策略与选择动作所用的策略 $\pi(s)$ 无关, 选择动作的策略直接决定了算法的效率[1-2]. 常用的动作策略选择方法中, 启发式策略选择主要研究如何利用先验或过程知识设计和优化启发式函数, 以提高 Q 学习的学习效率.

基于启发式策略选择的 Q 学习系统如图 1 所示. 该系统包含一个启发式策略学习模块, 此模块并不直接作用于值函数, 而是与 Q 学习系统结合起来调整搜索空间, 并不影响系统的收敛性[6-8]. 策略学习模块中建议动作的产生方法一般分为基于规则和基于启发式函数两类[7-8]. 前者以 if-then 的形式直接给出建议动作; 后者给每一个状态-动作对分配一个启发式函数 $H(s,a)$, 依据 $H(s,a)$ 间接地确定动作. 利用 $H(s,a)$ 产生建议动作的方式有多种, 但其本质是一样的, 本文采用下式得到建议动作 π_{ad}:

$$\pi_{ad}(s)=\max_{a\in A}H(s,a). \quad (2)$$

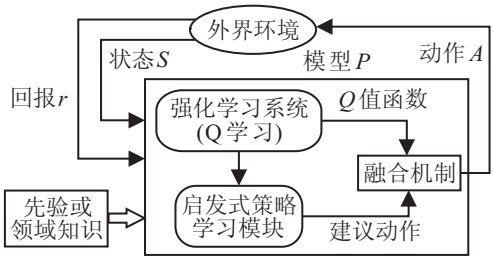


图 1 启发式 Q 学习原理

为了保证算法的收敛性和稳定性, 得到的建议动作采用下式的融合机制以概率形式作用于动作选择:

$$\pi(s)=\begin{cases}\pi_{ad}(s), \text{ rand}<\beta;\\ \pi_Q(s), \text{ otherwise.}\end{cases} \quad (3)$$

其中: rand 为在状态 s 时生成的随机数, β 为使用建议策略的概率 ($0\leqslant\beta\leqslant 1$). 当生成的随机数小于 β 时, 建议策略 π_{ad} 被采纳, 否则仍然根据 Q 值决定策略. β 在学习的过程中逐渐减小并最终趋于零, 有

$$\beta_t=\begin{cases}0, \ t<b;\\ \beta_{\text{int}}, \ b\leqslant t\leqslant c;\\ 0.99\beta_{t-1}, \ t>c.\end{cases} \quad (4)$$

其中: t 为学习幕数, b 为使用建议策略的开始幕数, c 为 β 逐渐减小的起始幕数.

2 基于拉普拉斯特征映射的启发式 Q 学习

2.1 强化学习中的拉普拉斯特征映射

拉普拉斯特征映射法是 Belkin 等[5,9-12]提出的一种计算效率较高的流形学习算法, 可以作为距离度量方法用于计算流形上的距离, 其基本思想是在高维空间中距离较近的样本点投影到低维目标空间中仍然保持邻近. LE 是一种典型的局部拓扑保持的降维方法, 通过极小化目标函数得到低维嵌入坐标, 并巧妙地将优化问题转换为求解矩阵的特征值和特征向量[9-12].

假设 X 是样本大小为 N 的数据集, 观测维数为 D, 内在维数为 n, 拟使用组合拉普拉斯算子, 则 LE 的

主要步骤如下[9-12].

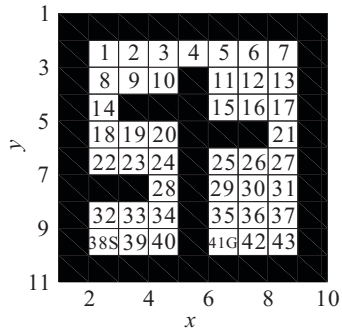
Step 1: 构造近邻图 G. 在数据集 X 中, 计算每个样本点 x_i 与其他样本点之间的欧氏距离 $d_E(x_i, x_j)$, 并利用 ξ-近邻 ($d_E(x_i, x_j) \leqslant \xi$ 的 x_i 与 x_j 有连接边) 或 KNN 相邻准则 (根据 $d_E(x_i, x_j)$, 与 x_i 最近的 K 个 x_j 有连接边) 构造无向近邻图 G.

Step 2: 在近邻图中, 为每条边设定一个权值 w_{ij}, 从而得到权值矩阵 W. 权值的选择有两种方式:

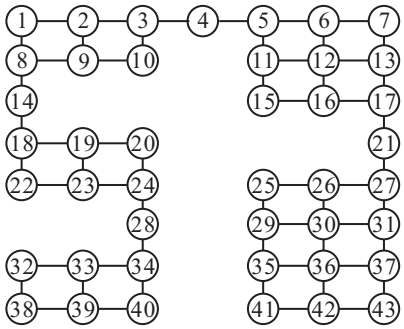
1) 热核法. 若点 x_i 与 x_j 是邻接的, 则边的权值为 $w_{ij} = \exp(-|x_i - x_j|^2/t)$, t 为比例参数, 否则 $w_{ij} = 0$.

2) 简单法. 若点 x_i 与 x_j 是邻接的, 则边的权值为 $w_{ij} = 1$, 否则 $w_{ij} = 0$.

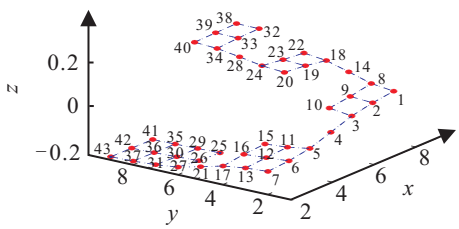
Step 3: 计算 n 维嵌入. 利用 $L = D - W$ 得到组合拉普拉斯矩阵 L, 其中度矩阵 D 为对角矩阵, $D(u, u) = \sum_{u \sim v} w_{uv}$, $u \sim v$ 为相邻节点. 计算 L 的 n 个最小特征值对应的特征向量 $f_1, f_2, \cdots, f_n$, 数据集 X 的低维嵌入可以表示为 $Y = [f_1, f_2, \cdots, f_n]^{\mathrm{T}}$.



(a) 格子世界地图



(b) 状态连接图



(c) SPVF

图 2　五房间格子世界的图描述

在强化学习问题中, 若通过采样获得了状态空间的连接关系, 则可以使用 LE 分析其内在拓扑结构, 并对状态空间进行降维和流形展开[5,9-12]. 例如, 状态空间在欧氏空间内不连续的五房间格子世界中 (如图 2(a) 所示), 其连接关系可用图 2(b) 表示, 相邻边权值均设为 1. 将 LE 应用该图时, 所得组合拉普拉斯矩阵的非零最小特征值对应的 Fiedler 特征向量如图 2(c) 所示. 图 2(c) 中, Z 轴为各状态相应的特征向量取值, 数字表示状态编号. 上述例子的状态空间是二维的, 形成的流形属于非封闭的无环类, 根据流形理论, 其可以在一维欧氏空间内展开[5,9-12]. 所以, Fiedler 特征向量可将图 2(a) 中二维欧氏空间内不连续的状态空间在一维欧氏空间内有效展开 (图 2(c) 的 Z 轴上). 由图 2 可见, 尽管状态 38 和状态 41 实际的图上距离最大 (也称流形距离最大), 但图 2(a) 中的欧氏距离并未反映出此关系. 而经过映射后, 图 2(c) Z 轴上的欧氏距离最大, 与图中距离一致. 其他状态与之类似[5,10].

在强化学习中, 将谱图理论应用于状态空间连接图所得的特征向量称为原型值函数 (PVF), 最小非零特征值对应的 Fiedler 特征向量称为 SPVF (Second PVF), 较小非零最小特征值对应的 PVF 称为低频 PVF[5,10-11]. 为了统一表述, 将相应的特征向量均称为 PVF.

2.2　基于拉普拉斯特征映射的启发式 Q 学习

在基于目标位置的强化学习任务中, 常用如下距离差作为启发式函数用于动作选择:

$$H(s, a) = d(s, s_g) - d(s', s_g). \tag{5}$$

其中: d 为某类距离, s_g 为目标状态. 如果 d 定义得当, 则能使各状态到目标位置的距离与值函数有类似的结构, 将其作为启发式函数可使向目标位置“靠近”的动作被采纳为建议动作 (见式 (2)). 对于状态空间在欧氏空间内不连续的任务, 欧氏距离不适宜作为启发式函数, 此时可以使用 LE 方法, 在特征映射空间里求取近似的流形距离用于设计启发式函数, 即

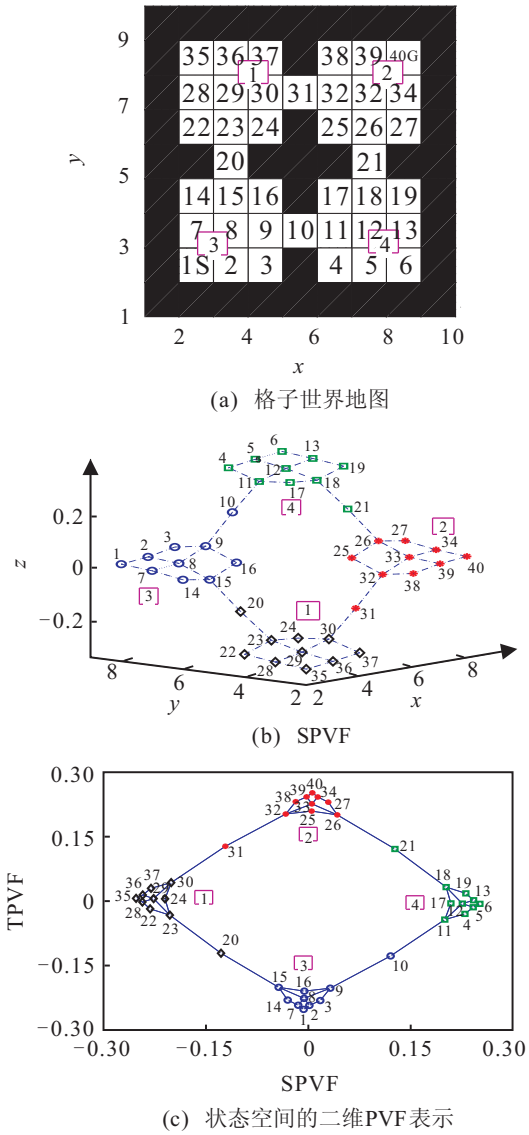
$$d_{\mathrm{pvf}}(s, a) = \sqrt{\sum_{i=2}^{n+1} (f_i(s_g) - f_i(s))^2}. \tag{6}$$

其中: f 为 PVF, n 为选择的 PVF 数目, 即状态空间所在流形的内在维数.

LE 作为一类基于局部拓扑保持的流形学习方法, 对流形上距离较远的点未作约束, 因此只能通过邻域的交织重叠展开流形. 这导致在 PVF 映射后, 各点间的欧氏距离并不一定能准确逼近流形上的距离. 但是, 在启发式函数的设计中, 低维欧氏空间中各映射点的距离只需保持与流形上距离相同的大小关系即可, 不需要精确逼近. 同时, 即使这种逼近的关系在

局部有出入, 但多数状态也能正确保持, 能从概率上保证多数向目标位置靠近的动作被采纳为建议动作, 从而提高算法的效率, 这正是欧氏距离在很多状态空间在欧氏空间上不连续的任务中作为启发式函数仍然有效的原因.

在流形学习中, LE对于有环的流形会失效. 同理, 在强化学习中, 若任务中子图间存在环, 则LE映射后的逼近距离不再与流形距离有同样的结构[5,9-12]. 此时, 可以采用升维映射法辅助解决, 即使用比状态空间内在维数高一维的特征映射. 例如, 在图3(a)所示的对称四房间格子世界中, 房间2与房间3状态间的实际距离较大, SPVF距离则较小(见图3(b) Z 轴方向的取值). 对于这种情况, 将状态空间映射到二维空间, 利用该二维空间形成的距离来设计启发式函数效果会更好(见图3(c), 其中TPVF表示第3个PVF, 即Third PVF). 需要说明的是, 升维映射法在理论上并不能保证对流形展开有好的效果, 所以是一种启发式的处理方法.



(a) 格子世界地图

(b) SPVF

(c) 状态空间的二维PVF表示

图 3 对称四房间格子世界的图描述

3 算法步骤、计算复杂度和适用范围

3.1 算法主要步骤

基于上述分析, 将流形学习中计算效率最高的LE方法用于Q学习, 可以得到一类新的启发式策略选择方法. 假设状态空间的观测维数为 D, 内在维数为 n, 新的启发式Q学习的步骤如下.

Step 1: Agent在任务环境中随机游走建立样本集. 利用第2.1节所述的方法选用 N 个样本建立环境状态的图论描述 $G=(V,W)$. 其中: V 为顶点集合, W 为邻接矩阵.

Step 2: 由式 $L=D-W$ 得到组合拉普拉斯算子并计算相应特征值, 根据任务的特点选择拉普拉斯特征映射的维数 n, 并利用Lanczos法求取 n 个或 $n+1$ 个最小特征值对应的特征向量.

Step 3: 对于子图间不存在环的任务, 在 n 维映射空间内利用式(6)计算各点到目标映射点的欧氏距离. 对于子图间存在环的任务, 需要在 $n+1$ 维空间内进行上述操作, 之后将所求距离带入式(5)求取启发式函数.

Step 4: 将所求的启发式函数用于Q学习中, 利用式(2)～(4)得到相应的启发式策略, 同时调节其融合参数.

3.2 计算复杂度

计算复杂度主要集中在建图、构造权值矩阵、求取特征向量和计算启发式函数3部分. 对于状态为连续的任务, 建图的计算复杂度为 $O(DN^2)$, 构造权值矩阵的复杂度最多不超过 $O(kDN)$ (k 为KNN中的参数)[12]. 对于离散环境的情况, 建图和构造权值的总计算复杂度为 $O(qDN)$, q 为状态的平均连接度. 求取启发式函数的计算复杂度为 $O(nN)$.

理论上讲, 对于有 N 个顶点的图 G, 其特征向量的计算复杂度为 $O(N^3)$, 但是强化学习的状态连接图一般较为稀疏, 所以计算复杂度大幅降低[9-12]. 求取稀疏的拉普拉斯矩阵 n 个特征向量的计算复杂度为 $O(npN^2)$, p 为图的稀疏程度, 即拉普拉斯矩阵中非零元素的比例, N 越大, p 越小[12]. 在HARL中, 首先也需要建图, 然后采用基于动态规划的启发式信息反向传播技术求取建议策略的计算复杂度 $O(DN^2)$. 常用的求取图上最短路径算法的计算复杂度一般为 $O(DN^2)$.

由上述算法步骤和计算复杂度可知, 建图的精度直接影响后续降维和流形展开的质量. 精度越高, 后续表示的效果越好, 但是计算复杂度也会快速增加, 同时LE方法在特征值求取部分难以实现增量计算. 所以, 从计算复杂度角度而言, 基于LE的启发式函数设计方法复杂度较大, 但相对于HARL和求取最短路

径的算法仍然有所提升, 并且 D/n 越大, 提升效率越明显. 同时, 所求得的特征向量还可以用于降维、任务分解和值函数泛化等, 若将多种方法结合起来则其优势较为明显.

3.3　适用范围

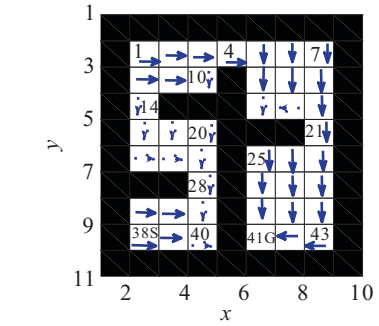
所提出的方法使用了基于谱图理论的LE. LE采样建图时有一个对称化的操作, 即得到的拉普拉斯矩阵式是自然对称的[9,12]. 对于状态连接关系为有向图的任务, 上述的对称化操作实质上将其变为无向图, 会使低维PVF空间内各映射点间的欧氏距离不能正确反映流形上实际距离的大小关系, 这也是第4节仿真实验只用格子世界而未使用常见的倒立摆和小车爬山等测试例子的原因. 距离扭曲的情形会出现在流形内在维数估计错误时, 流形学习对于有环的流形可能会失效.

综上所述, 本文所提出算法主要适用于状态空间内在流形维数能较好估计、流形结构不存在环且连接关系为无向图的任务. 对于流形结构存在环的任务, 使用升维映射会使多数状态间的欧氏距离与流形距离较一致, 从概率上讲仍然能够提高算法性能.

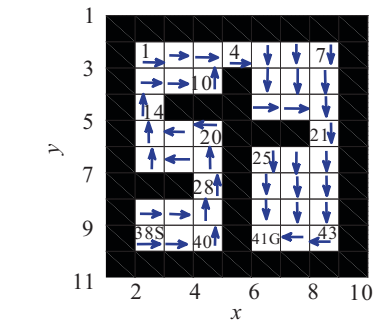
4　仿真实验和结果分析

为了验证新方法的有效性, 分别在子任务间存在环与不存在环的两类格子世界中进行仿真. 仿真中, 格子世界的状态是离散的, Agent有向东、向西、向南、向北4个确定性动作, 碰墙后状态不变, 立即回报值函数在非目标状态值为0, 在目标状态为1. 仿真中算法均采用Q学习, 其中 $\alpha = 0.01$, $\gamma = 0.98$, 每组实验各算法均独立运行20次. 邻接矩阵 W 的设定方式采用第3.1节描述的简单法.

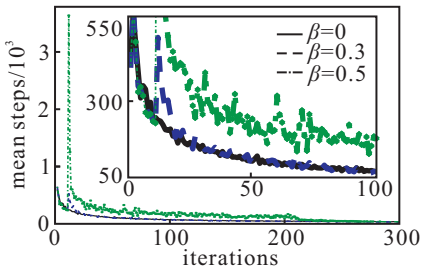
在五房间格子地图中, 目标状态设为41, Agent的初始状态在38 (见图2(a)). 由于该任务的子图间不存在环, 只使用单个SPVF即可, 图4为相应的仿真结果, 图4(a)为使用欧氏距离得到的最好参考策略, 虚线箭头部分的参考策略并非最优策略. 计算欧氏距离时, 采用图2(a)所示的坐标, 例如状态1的坐标为(2, 2), 状态41的坐标为(7,10). 图4(b)为所提方法使用SPVF距离获得的参考策略, 其均为最优策略. 图4(c)对比了欧氏距离作为启发式函数时不同动作融合概率 β 的学习曲线, 图中小图为前多幕学习曲线的放大. β 在10～200幕期间一直为初始值, 从200幕开始以式(5)方式逐渐减小. 由图4(c)对比结果可知, 欧氏距离作为启发式函数在该任务中是失败的, 不仅未提高Q学习的效率, 反而更糟, β 越大, 效率越低, 原因在于该任务的状态空间在欧氏空间内不连续. 图4(d)为使用SPVF距离作为启发式函数的相应结果, 可见SPVF距离能够显著提高算法效率.



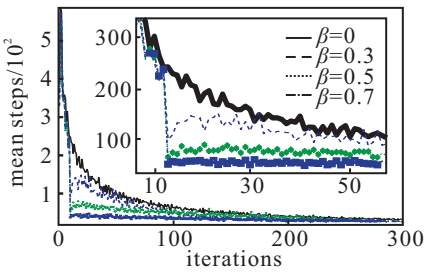
(a) 欧氏距离的策略



(b) SPVF距离的策略



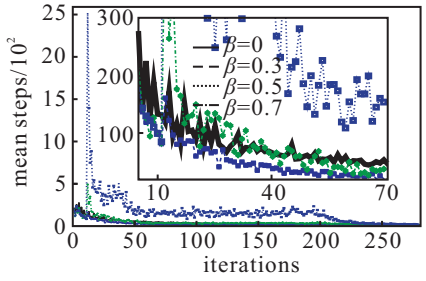
(c) 欧氏距离仿真结果



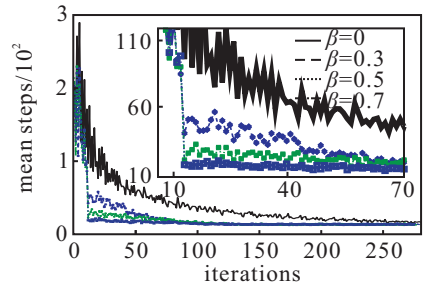
(d) SPVF距离仿真结果

图 4　五房间格子世界的仿真结果

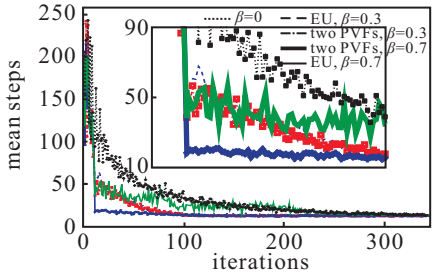
在对称四房间地图中, 由于各房间的连接关系存在环, 使用二维特征映射的距离, 即使用SPVF和TPVF. 任务状态的起始位置为状态1, 目标位置为40 (见图3(a)), 其相关参数与五房间格子世界任务一致, 图5为相应的仿真结果. 由图5(a)和图5(b)可见, 仅使用SPVF距离降低了算法的效率, 使用两个PVF距离提高了算法效率. 图5(c)对比了使用两个PVF距离与欧式距离的仿真结果, 其中EU简写表示欧氏距离. 对比结果表明, 使用欧氏距离的算法性能劣于使用两个PVF距离的结果, 但与未使用启发式函数的Q学习(即 $\beta = 0$)比较仍有提高, 原因是此地图中障碍物相对稀疏, 欧氏距离较大程度上反映了实际的图上距离.



(a) SPVF距离仿真结果



(b) 二维PVF距离仿真结果



(c) 欧氏距离与二维PVF距离比较

图 5 对称的四房间格子世界的仿真结果

5 结 论

对于状态空间在欧氏空间内不连续、在流形上连续的任务, 利用欧氏距离作为启发式函数进行动作选择时效果不理想. 针对该问题, 将流形学习中计算复杂度较小的LE方法引入启发式Q学习中, 提出了一种基于谱图理论的启发式函数设计方法, 并分析了计算复杂度和优缺点. 格子世界的仿真结果验证了所提方法的有效性. 所提启发式函数设计方法只能用于状态空间的连接关系为无向图的任务, 且需要知道任务状态空间的全局连接关系和判断子图间是否存在环, 用于状态空间较大的任务时采样和计算复杂度较高. LE不仅能用于启发式函数的设计, 还可以用于值函数泛化和子任务分解. 下一步工作将研究如何在PVF框架下将多种方法结合使用, 间接地降低计算复杂度, 同时将算法扩展到有向图任务中.

参考文献(References)

[1] Sutton R S, Barto A G. Reinforcement learning: An Introduction[M]. Cambridge: MIT Press, 1998:1-5.

[2] 高阳, 陈世富, 陆鑫. 强化学习研究综述[J]. 自动化学报, 2004, 30(1): 86-100.
(Gao Y, Chen S F, Lu X. Research on reinforcement learning technology: A review[J]. Acta Automatica Sinica, 2004, 30(1): 86-100.)

[3] 吴军, 徐昕, 王健, 等. 面向多机器人系统的增强学习研究进展综述[J]. 控制与决策, 2011, 26(11): 1601-1610.
(Wu J, Xu X, Wang J, et al. Recent advances of reinforcement learning in multi-robot systems: A survey[J]. Control and Decision, 2011, 26(11): 1601-1610.)

[4] 陈宗海, 杨志华, 王海波, 等. 从知识的表达和运用综述强化学习研究[J]. 控制与决策, 2008, 23(9): 962-968.
(Chen Z H, Yang Z H, Wang H B, et al. Overview of reinforcement learning from knowledge expression and handling[J]. Control and Decision, 2008, 23(9): 962-968.)

[5] 朱美强, 李明, 张倩. 一类用于井下路径规划问题的Dyna_Q学习算法[J]. 工矿自动化, 2012(12): 71-75.
(Zhu M Q, Li M, Zhang Q. A dyna Q-learning algorithm in underground path planning[J]. Industral and Mine Automation, 2012(12): 71-75.)

[6] Bianchi R A C, Ribeiro C H C, Costa A H R. Accelerating autonomous learning by using heuristic selection of actions[J]. J of Heuristics, 2008, 14(2): 135-168.

[7] Marek G. Improving exploration in reinforcement learning through domain knowledge and parameter analysis[D]. York: Department of Computer Science, University of York, 2010: 34-36.

[8] Bradley K W, Peter S. Augmenting reinforcement learning with human feedback[C]. The 28th ICML Workshop on New Developments in Imitation Learning. Washington, 2011: 127091.

[9] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation[J]. Neural Computation, 2003, 15(6): 1373-1396.

[10] 朱美强, 程玉虎, 李明, 等. 一类基于谱方法的强化学习混合迁移算法[J]. 自动化学报, 2012, 38(11): 1765-1776.
(Zhu M Q, Cheng Y H, Li M, et al. A hybrid transfer algorithm for reinforcement learning based on spectral method[J]. Acta Automatica Sinica, 2012, 38(11): 1765-1776.)

[11] Mahadevan S. Learning representation and control in Markov decision processes: New frontiers[J]. Foundations and Trends in Machine Learning, 2009, 4: 403-565.

[12] 曾宪华. 流形学习的谱方法相关问题研究[D]. 北京: 北京交通大学计算机与信息学院, 2009: 21-23.
(Zeng X H. Study on several issues of spectral method for manifold learning[D]. Beijing: School of Compuetr and Information Technology, Beijing Jiaotong University, 2009: 21-23.)

(责任编辑: 郑晓蕾)