

文章编号:1001-9081(2009)03-0871-03

基于三角剖分的小脑模型在增强学习中的应用

孙方义, 郑志强

(国防科学技术大学 机电工程与自动化学院, 长沙 410073)

(sunfangyi@nudt.edu.cn)

摘要: 研究了一种基于三角剖分的小脑模型的增强学习控制器设计方法, 并应用于机器人足球中单机器人截球的学习控制中。该方法通过在 Markov 决策过程状态空间中引入基于单纯形的库恩三角化, 实现基于三角剖分的线性值函数逼近, 从而有效提高了增强学习控制器对连续状态空间马氏决策问题的泛化性能。针对机器人截球学习控制的仿真研究表明, 采用基于三角剖分的小脑模型进行值函数逼近的增强学习控制器能够获得优于已有基于均匀编码的小脑模型方法的学习效率和泛化性能。

关键词: 增强学习; 小脑模型关节控制器; 库恩三角化; Markov 决策过程

中图分类号: TP181 文献标志码:A

CMAC application using triangulation in reinforcement learning

SUN Fang-yi, ZHENG Zhi-qiang

(College of Mechatronics and Automation, National University of Defense Technology, Changsha Hunan 410073, China)

Abstract: A reinforcement learning controller based on CMAC neural networks using triangulation was studied and applied to the learning control of intercepting a ball in the RoboCup. By utilizing Kuhn triangulation based on simplex interpolation in the continuous state space of Markov Decision Processes (MDPs), the value functions of MDPs were approximated with linear triangulation so that the generalization ability of the CMAC-based reinforcement learning controller could be improved. Simulation results on the learning control of intercepting a ball show that the CMAC-based learning controller using triangulation is much more efficient than the learning controller based on CMAC uniform coding.

Key words: reinforcement learning; Cerebellar Model Articulation Controller (CMAC); Kuhn triangulation; Markov decision process

0 引言

小脑模型关节控制器 (Cerebellar Model Articulation Controller, CMAC) 是 J. S. Albus 于 1975 年首次提出的用于模拟小脑功能的神经网络模型^[1]。由于它是基于局部学习的神经网络, 对每一个输入向量只有一小部分神经元产生响应, 因此具有学习速度快、泛化能力强、易于实现等特点, 已被成功地应用于复杂函数的近似、系统辨识和控制等领域中。由于 CMAC 具有较强的函数逼近和泛化能力, 因此它在增强学习^[2]领域的研究中也得到了普遍重视。

然而由于 CMAC 最初的接收域函数是零阶的 (即为矩形接收域), 使得网络映射空间表达易出现不连续, 特别是在输入维数较高时, 需要大量的存储空间, 降低了学习效率。徐昕在 2002 年提出了一种基于变尺度编码 CMAC 神经网络的增强学习控制器设计方法^[3], 在该方法中通过对状态空间变尺度重叠量化编码, 实现了基于 CMAC 的多分辨率值函数逼近。但是由于该方法需要先验知识, 因此并不是一种更为一般化的方法, 并且在状态空间维数增大时, 仍然面临着记忆单元急剧增大的问题。

将多维空间线性插值^[4]与 CMAC 结合是解决维数增大时所带来的问题的重要途径, 本文在保留 CMAC 原有增强和局部特性的基础上, 采用多维空间中的单纯形^[5]实现对状态空间的线性插值, 提出一种基于三角剖分的 CMAC 神经网络 (Triangulation-CMAC, TRI-CMAC), 并以 RoboCup 中的截球为例和传统的基于表格的 CMAC (Tabular-CMAC, TAB-CMAC)

进行了对比。结果表明, 基于三角剖分的 CMAC 能够有效地利用状态空间相邻区域的泛化信息, 显著地提高了算法的学习效率, 获得优于已有算法的性能。

1 CMAC 神经网络模型

CMAC 是一个具有三层映射结构的前馈神经网络, 如图 1 所示, 在 CMAC 网络中, 第一层映射是对输入状态空间 S 的层叠式量化编码, 即采用多个具有不同偏移量的量化网络将输入空间映射到一个高维的离散编码空间。设 CMAC 的量化网格数目为 C (C 通常又称为泛化参数), 即 A_1, A_2, \dots, A_c , 输入状态维数为 n , 每个输入的量化区间个数为 q , 则每个网格的离散单元数目为 q^n 。对于输入状态空间的一个点 S , 在每个网格中将有唯一的单元被激活, 即取值为 1, 其余单元为 0。在整个层叠量化编码结构中共有 C 个激活单元与一个输入状态对应。第二层映射是将高维的离散编码空间映射到一个一维的物理地址空间, 若采用一对一的映射, 则物理地址空间的单元总数为 $C \times q^n$ 。通常为减少存储量, 在第二层映射中采用 Hash 技术。设 $F(s) = (f_1, f_2, \dots, f_N)$ 表示对应物理地址空间的特征向量, N 为物理地址空间的单元数, 对于激活单元, $f_i = 1$, 否则 $f_i = 0$ 。 f_i 的激活与否, 由第一层的层叠式量化编码结构决定, 因此对应一个输入状态, 将同时有 c 个物理地址单元被激活。

CMAC 网络的最后一层映射是输出映射, 为输出层权值的线性加权组合。设 $W = (w_1, w_2, \dots, w_N)$ 为对于物理地址空间的权值向量, 则 CMAC 网络的输出计算公式为:

收稿日期: 2008-09-09; 修回日期: 2008-10-31。

作者简介: 孙方义(1985-), 男, 吉林桦甸人, 硕士研究生, 主要研究方向: 机器人智能控制; 郑志强(1965-), 男, 湖南常德人, 教授, 博士生导师, 主要研究方向: 机器人控制、多机器人协作、导航与制导控制。

$$Y(s) = \mathbf{W}^T F(s) = \sum_{i=1}^N w_i f_i \quad (1)$$

对于通常采用的均匀编码 CMAC 神经网络,在给定泛化参数 C 和每个输入变量的量化区间个数 q 后,CMAC 网络的结构就能够完全确定。

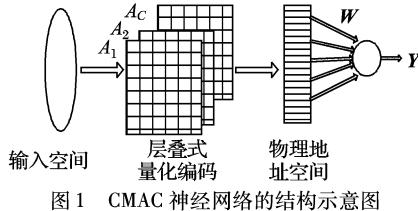


图 1 CMAC 神经网络的结构示意图

2 基于 TRI-CMAC 的增强学习控制器

CMAC 作为一种局部泛化神经网络,除了在监督学习中得到成功应用外,在增强学习中也广泛地被用于求解大规模和连续状态空间 Markov 决策过程的学习控制问题^[6]。在基于 CMAC 神经网络的增强学习算法中,CMAC 网络通常作为 Markov 决策过程的行为值函数逼近器,学习控制器根据 CMAC 网络的输出来确定控制量输出。

Markov 决策过程可以用五元组 $\{S, A, R, P, J\}$ 来描述,其中 S 为状态空间,通常为连续集合, A 为具有 l 个元素的离散行为空间, R 为回报函数, P 为状态转移概率, J 为优化性能指标,通常为各时刻回报的折扣加权和。我们采用行为值函数 $Q(s, a_i)$ ($i = 1, 2, \dots, l$) 来逼近 Markov 决策过程 (Markov Decision Process, MDP) 的 l 个离散行为,其定义为:

$$Q^\pi(s, a_i) = E^\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a_i \right] \quad (2)$$

其中, π 为平稳行为策略。对应最优性能指标 J^* 的最优值函数 $Q^*(s, a)$ 和最优策略 π^* 分别为:

$$Q^*(s, a_i) = \max_{\pi} E^\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a_i \right] \quad (3)$$

$$\pi^*(s) = \max_a Q^*(s, a) \quad (4)$$

基于 CMAC 的增强学习控制器通过如下的 ε 贪心策略确定控制量输出:

$$a_t = \begin{cases} \operatorname{argmax} Q(s, a_i), & \text{以概率 } 1 - \varepsilon \\ a_{\text{rand}}, & \text{以概率 } \varepsilon \end{cases} \quad (5)$$

利用 CMAC 实现对具有大规模和连续状态空间的学习,是基于 CMAC 能够对相似的输入产生相似的输出。由于 CMAC 的局部泛化是通过对 MDP 状态空间的量化编码实现的,在面对大规模和连续状态空间时,虽然可以通过增加量化等级数来提高状态空间的分辨率,但这会导致算法的计算量和存储量迅速增加,同时由于 CMAC 最初的接收域是矩形区域,网络映射空间是不连续的,因此不利于算法的泛化性能。如何在不增加量化等级数目的条件下,提高 CMAC 网络对于高维连续空间的泛化性能是设计基于 CMAC 的增强学习控制器需要解决的一个重要问题。

本文采用一种基于三角剖分的线性插值方法来提高 CMAC 神经网络在求解 Markov 决策问题时的泛化性能。在 CMAC 网络的第一层映射中,引入基于三角剖分的多维线性插值(又称为库恩三角化,Kuhn Triangulation)^[7],如图 2 所示,在状态空间是 3 维时,每一个单元立方体被分割成 6 个棱锥体,同理,如果状态空间为 d 维时,每一个单元超立方体被分割成 $d!$ 个单纯形。对于单纯形内的某一个点,其值可以看成是所在单纯形的 $(d+1)$ 个顶点的值的加权和。

假定状态空间的每一个量化单元为单位超立方体,其中一个顶点为 $(x_0, x_1, \dots, x_{d-1}) = (0, 0, \dots, 0)$,其所在对角线

的另一端为 $(1, 1, \dots, 1)$ 。因此,超立方体内的每一个单纯形可以看成为 $(0, 1, \dots, d-1)$ 的一种排列。我们按如下步骤实现基于库恩三角化的线性插值:

1) 首先将所求的点 $(x_0, x_1, \dots, x_{d-1})$ 归一化为 $(x'_0, x'_1, \dots, x'_{d-1})$ 。

2) 对 $(x'_0, x'_1, \dots, x'_{d-1})$ 中的元素进行排序,则 $1 \geq x_{j_0} \geq x_{j_1} \geq \dots \geq x_{j_{(d-1)}} \geq 0$,因此 $(j_0, j_1, \dots, j_{(d-1)})$ 表示了该点所在的单纯形位于超立方体中的位置,设该单纯形的 $(d+1)$ 个顶点分别为 $\{\varphi_0, \varphi_1, \dots, \varphi_d\}$ 。

3) 定义所求点的重心坐标 $\lambda_0, \lambda_1, \dots, \lambda_d$,其满足 $\sum_{k=0}^d \lambda_k = 1, \lambda_0 = 1 - x_{j_0}, \lambda_1 = x_{j_0} - x_{j_1}, \dots, \lambda_k = x_{j_{(k-1)}} - x_{j_k}, \dots, \lambda_d = x_{j_{(d-1)}} - 0 = x_{j_{(d-1)}} \circ$

因此所求点的值为 $x = \sum_{k=0}^d \lambda_k \varphi_k$ 。

通过在 CMAC 神经网络的第一层映射中引入库恩三角化,其最终的输出为:

$$Y(s) = \sum_{i=1}^N \left(\sum_{k=0}^d \lambda_k \varphi_k \right) f_i \quad (6)$$

对于状态空间内的任意一点,都能找到相应的单纯形并通过上式计算出最终的输出值。

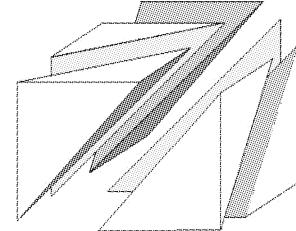


图 2 三维空间的库恩三角化

基于 TRI-CMAC 和 Sarsa(λ)学习算法^[8]的增强学习控制器设计和算法流程如下:

1) 初始化 TRI-CMAC 神经网络。根据行为集合的元素个数 l 确定 l 个行为值函数 $Q(s, a_i)$,并在确定 TRI-CMAC 网络的量化编码结构中引入库恩三角化,确定泛化参数 C 和初始化 TRI-CMAC 网络的权值向量 $\mathbf{W} = (w_1, w_2, \dots, w_N)$;

2) 初始化有关学习参数 $\alpha, \gamma, \varepsilon$,设置权值的适合度轨迹 $e^t = (0, 0, \dots, 0)$,学习周期数 Episode 为 0;

3) 初始化被控对象的状态,采样时刻 $t = 0$;

4) 对当前状态 s_t ,按照行为选择策略(5)确定当前控制量 a_t ;

5) 观测回报 $r(s_t, a_t)$ 和下一时刻状态 s_{t+1} ,并且按照行为选择策略(5)确定控制量 a_{t+1} ;

6) 根据状态 s_t ,按照式(6)确定与行为 a_t 对应的行为值函数 $Q(s_t, a_t)$,同理确定 $Q(s_{t+1}, a_{t+1})$,并根据状态 s_t 在 TRI-CMAC 网络中被激活的 C 个单纯形所对应的 $C \times (d+1)$ 个物理地址单元构成的集合为 A ,根据如下公式对每个 TRI-CMAC 网络的权值进行迭代:

$$e_j^i = \begin{cases} \lambda_j^k, & f_j^i \in A^i \\ \gamma e_j^i, & \text{其他} \end{cases} \quad (7)$$

$$w_j^i = w_j^i + \alpha_t(r(s_t, a_t) + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)) e_j^i \quad (8)$$

λ_j^k 为被激活的单纯形对应的第 k 个顶点相应的重心坐标分量。

7) 若 s_{t+1} 为终端状态,则当前学习周期结束,并判断是否满足算法停止条件,若满足,则算法结束,否则 Episode = Episode + 1,返回 2);否则 $t = t + 1$,返回 5)。

下面我们分析一下该算法的计算复杂度。假设对于给定

问题的 d 维状态空间,其行为集合元素个数为 l ,经过算法第 1)步的初始化,存在 N 个特征值,适合度轨迹 e' 含有 m 个非零元素,通常情况下, $m \ll N$ 。第 5)~7) 为该算法的主循环,其中第 5) 步的计算复杂度为 $O(l)$; 第 6) 步的计算复杂度为 $O(m(d+1)) + O(m)$,前一项为计算 $Q(s_i, a_i)$ 的计算复杂度,后一项为权值更新的计算复杂度; 第 7) 步计算复杂度为 $O(1)$,因此整个算法的计算复杂度为 $O(dm)$ 。而一般的基于均匀编码 CMAC 的 Sarsa(λ) 算法的计算复杂度为 $O(m)$,对于给定问题的 d 维状态空间,由于 $m \ll N$,因此改进算法的计算复杂度是完全可以接受的。

3 基于 TRI-CMAC 的截球学习仿真

本章我们将用上述算法讨论机器人足球中的截球问题。如图 3 所示,机器人试图截住正在向前滚动的足球,为此机器人需要运动到合适的一点上才能够成功。我们采用本实验室开发的 RoboCup 中型组仿真来做这个实验。在该仿真平台中,机器人能够获得自身与足球的相关状态信息,坐标系采用以机器人为中心的相对坐标系,其正前方为 x 轴正方向,采用右手坐标系。下面将以 Markov 决策模型描述该截球问题。

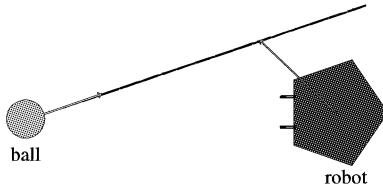


图 3 截球示意图

该截球问题的状态空间如下:

- 1) 足球相对机器人的位置(二维: (p_x, p_y));
- 2) 足球相对机器人的速度(二维: (v_x, v_y));
- 3) 足球自身速度的大小(一维: v_b)。

在该问题中,状态空间维数为 5,我们分别采用 CMAC 和 TRI-CMAC 网络作为线性函数估计器。在 CMAC 网络中,状态空间的每一维进行均匀量化,而在 TRI-CMAC 网络中,则是在上述量化的基础上,引入库恩三角化,即基于单纯形的线性插值。各个输入变量的量化区间参数为: $p_x: (-100, 600)$, $q_0 = 10$; $p_y: (-350, 350)$, $q_1 = 10$; $v_x: (-300, 300)$, $q_2 = 5$; $v_y: (-300, 300)$, $q_3 = 5$; $v_b: (0, 400)$, $q_4 = 5$ 。

其中泛化参数 $C = 16$,可知物理地址单元总数为: $N = q_0 \times q_1 \times q_2 \times q_3 \times q_4 \times C = 200\,000$ 。

行为空间采用离散化的取值,即机器人向 8 个间隔为 $\pi/4$ 弧度的方向以最大的加速度运动, $A = \{(ang, acc) | (0, 300), (\pi/4, 300), (\pi/2, 300), (3\pi/4, 300), (\pi, 300), (5\pi/4, 300), (3\pi/2, 300), (7\pi/4, 300)\}$,其中 ang 为机器人运动的方向, acc 为机器人运动的加速度。

在截球学习的过程中,当球在机器人正前方一定范围以内时便认为机器人成功地截住了球,并获得一个正的奖赏值。否则的话,将会获得一个负的惩罚值。其他有关的学习参数为: $\alpha = 0.5$, $\gamma = 0.9$, $v = 0.2$ 。

由于是大规模状态空间,为减少权值存储量,在 CMAC 和 TRI-CMAC 网络中采用了 Hash 映射技术。在该机器人足球仿真平台中,截球学习是以一个接一个的学习周期方式进行的,当足球被成功截住或拦截失败时,结束当前的学习周期并开始下一个学习周期。仿真实验结果如图 4 和 5 所示。其中图 4 为基于 CMAC 网络的截球学习曲线,图 5 为基于 TRI-CMAC 网络的截球学习曲线。从图中可以看出,本文提出的基于三角剖分的小脑模型增强学习控制器无论是在学习的成功率上,还是在学习成功时所需要的步数上,都要远优于传统

的均匀编码的小脑模型增强学习控制器。

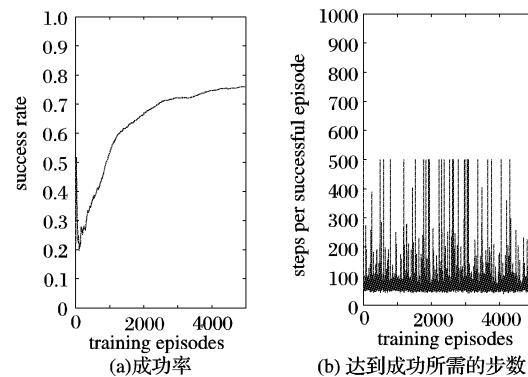


图 4 基于 CMAC 网络增强学习

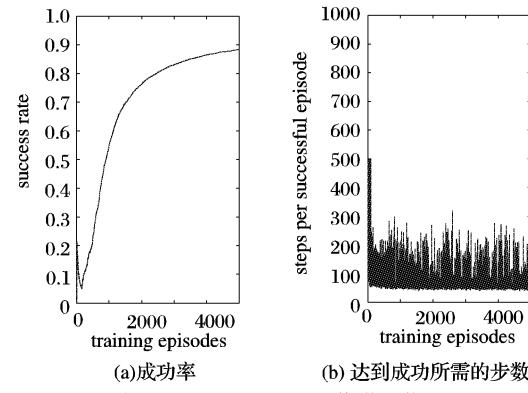


图 5 基于 TRI-CMAC 网络增强学习

4 结语

对于大规模连续状态空间 Markov 决策过程的学习控制问题,本文研究了一种基于三角剖分的小脑模型增强学习控制器设计方法。该方法通过对 CMAC 网络中的量化网格引入库恩三角化,在每个网格中实现基于单纯形的多维线性插值,以实现对 MDP 值函数的连续逼近,从而提高小脑模型在连续状态空间增强学习问题中的泛化性能。在单机器人截球学习控制问题的仿真研究中,上述方法获得了远优于已有方法的学习效率。本文的研究结果对于解决大规模连续状态空间的增强学习问题具有重要的意义。

参考文献:

- [1] ALBUS J S. A new approach to manipulator control: The cerebellar model articulation controller(CMAC) [J]. Journal of Dynamic Systems, Measurement, and Control, 1975, 97(3): 220~227.
- [2] SUTTON R S, BARTO A G. Reinforcement learning: An introduction [M]. Cambridge, MA: MIT Press, 1998.
- [3] 徐昕, 贺汉根. 基于变尺度编码 CMAC 的增强学习控制器及其应用 [J]. 模式识别与人工智能, 2002, 15(3): 264~269.
- [4] DAVIES S. Multidimensional triangulation and interpolation for reinforcement learning[C]// Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 1997, 9: 1005~1011.
- [5] MUNOS R, MOORE A. Variable resolution discretization for high-accuracy solutions of optimal control problems[C]// International Joint Conference on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 1999: 1348~1355.
- [6] KAEELBLING L P, LITTMAN M L, MOORE A W. Reinforcement learning: A survey[J]. Journal of Artificial Intelligence Research, 1996, 4: 237~285.
- [7] PLAZA A. The eight-tetrahedra longest-edge partition and Kuhn triangulations[J]. Computers & Mathematics with Applications, 2007, 54(3): 427~433.
- [8] SINGH S P, SUTTON R. Reinforcement learning with replacing eligibility traces[J]. Machine Learning, 1996, 22(1/2/3): 123~158.