

XCSG 在多机器人强化学习中的应用

邵杰^{1,3} 杜丽娟² 杨静宇³

(郑州成功财经学院信息工程系 郑州 451200)¹ (商丘工学院信息与电子学院 商丘 476000)²

(南京理工大学计算机科学与技术学院 南京 210094)³

摘要 XCS 分类器在解决机器人强化学习方面已显示出较强的能力,但在多机器人领域仅局限于 MDP 环境,只能解决环境空间较小的学习问题。提出了 XCSG 来解决多机器人的强化学习问题。XCSG 建立低维的逼近函数,梯度下降技术利用在线知识建立稳定的逼近函数,使 Q -表格一直保持在稳定低维状态。逼近函数 Q 不仅所需的存储空间更小,而且允许机器人在线对已获得的知识进行归纳一般化。仿真实验表明,XCSG 算法很好地解决了多机器人学习空间大、学习速度慢、学习效果不确定等问题。

关键词 强化学习,多机器人,学习分类器,梯度下降法的学习分类器

中图法分类号 TP181 文献标识码 A

Applications of XCSG in Multi-robot Reinforcement Learning

SHAO Jie^{1,3} DU Li-juan² YANG Jing-yu³

(Zhengzhou Chenggong University of Finance and Economics, Department of Information Engineering, Zhengzhou 451200, China)¹

(School of Information and Electronic, Shangqiu Institute of Technology, Shangqiu 476000, China)²

(School of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, China)³

Abstract XCS classifier system has been shown to solve machine-learning problems in a competitive way. However, in multi-robot problems, XCS is restricted to solve very small problems modeled by a Markov decision process. In this paper a new learning technique XCSG that combines XCS and gradient descent methods was proposed to solve multi-robot machine-learning problems. XCSG builds low-dimensional approximation of the function, and gradient descent techniques use on-line knowledge to establish a stable approximation of functions, so that the Q -form has been maintained at a low-dimensional stable state. Approximate of the function not only requires smaller storage space, but also allows the robot online knowledge is summarized on the generalization. Simulation results show that XCSG algorithm solves the multi-robot reinforcement learning in a large space, slow learning, learning uncertainty and other issues.

Keywords Reinforcement learning, Multi-robot, Accuracy-based learning classifier system (XCS), Accuracy-based learning classifier system with gradient descent method(XCSG)

Q -学习作为一种典型的强化学习方法,不需要建立环境和任务的精确数学模型,已被广泛地应用于机器人领域^[1]。但在多机器人的学习过程中, Q -函数是通过每个机器人输入状态-动作对对应一个输出值的表格形式来表示的。对于小型的状态空间来说, Q -学习方法效果很好,但随着机器人状态-动作对的增多、学习空间的增大, Q -表格所要求的存储空间、学习收敛的时间以及每次迭代的时间都会迅速增加, Q -学习已无法解决复杂环境的学习,这也是 Q -学习在实际应用中的主要缺陷。

基于上述分析,本文提出了用 XCSG 来解决多机器人的强化学习问题。XCSG 建立低维的状态-动作对逼近函数 Q 和状态-动作对到环境回报值的映射关系。利用梯度下降技术在线建立的稳定逼近函数 Q ,可使状态-动作对到环境回报值的映射 Q -表格一直保持在低维状态。逼近函数 Q 所需的

存储空间更小,XCSG 算法允许机器人在线对已获得的知识进行归纳一般化。仿真实验表明,XCSG 算法很好地解决了多机器人学习空间大、学习速度慢、学习效果不确定等问题。

1 强化学习相关技术

机器人强化学习(Reinforcement Learning, RL)^[2]是通过 trial and error 试错操作感知环境状态和从环境中获得回报,最终获得最优行为策略的动态在线学习系统。在时刻 t ,机器人所处状态为 s_t ,机器人有可能对环境选择执行动作集 A 中的动作 a_t ,并从环境获得回报值 r_{t+1} ,机器人状态从状态 $s_t \rightarrow s_{t+1}$ 。机器人在 k 时刻获得的最大期望回报为:

$$Q(s, a) = E\left[\sum_{k=0}^{\infty} \gamma^k r_{t+1+k}\right] \quad (1)$$

式中, γ 为折扣因子($0 \leq \gamma \leq 1$),表示未来回报的重要程度。 γ 值越大,表示未来回报越重要。

到稿日期:2012-11-01 返修日期:2013-03-14 本文受国家自然科学基金(90820004)资助。

邵杰(1965—),男,博士,教授级高工,主要研究方向为多机器人系统等;杨静宇(1941—),男,教授,博士生导师,主要研究方向为模式识别、智能机器人等。

1.1 Q学习

Q学习是由 Watkins 1989 年提出的基于 Markov 环境的一种学习算法 (MDP)^[3]。其原理是通过对感知的环境状态采取各种试探动作,获得此种试探动作对此种环境状态的奖励或者惩罚信号,能不断调整学习策略以获得较大的奖励或较小的惩罚,同时反馈给其它机器人,并能够保证学习的收敛性。MDP 定义如下:

1) 设定状态空间 $S = \{s_1, s_2, \dots, s_n\}$, 在任意时刻 t 的状态 $s_t \in S$ 。

2) 依据机器人的状态 s_t 设立动作集合 $A(s_t)$, 任一时刻 t 的动作 $a_t \in A(s_t)$ 。

3) 状态转换函数 $T(s, a, s')$, 映射 $\langle s, a \rangle$ 对到后续状态 s' 的概率分布。

4) 回报函数 $R(s, a, s')$, 表示机器人采取动作 s' 从状态 s 到 s' 所获得的回报。机器人在时刻 t 获得的回报为 r_t 。

Q学习算法和优化策略迭代算法如下:

$$Q(s_t, a_t) = Q(s_t, a_t) + \beta(r_t + \gamma \max_{a'} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)) \quad (2)$$

$$\pi(s_t, a_t) = \arg \max_a Q(s_{t+1}, a) \quad (3)$$

式中, $\beta(0 \leq \beta \leq 1)$ 是学习率, $a \in A$, $\gamma(0 \leq \gamma \leq 1)$ 是折扣因子。

1.2 梯度下降法 Q学习

将梯度下降法^[4-6]应用于 Q学习, 我们是要最大程度地减少依 $r_{t+1} + \lambda \max Q(s_t, a)$ 估计得到的期望回报值与依 $Q(s_{t+1}, a_{t+1})$ 得到的实际回报间的误差。用梯度下降形成的权重矩阵 W 参数化 Q学习逼近函数。在每个时刻 $t+1$, 每个权重 w 的变化为 Δw , 此为基本梯度算法:

$$\Delta w = \beta(r_{t+1} + \gamma \max Q(s_{t+1}, a) - Q(s_t, a_t)) \frac{\partial Q(s_t, a_t)}{\partial w} \quad (4)$$

基本梯度算法用于 MDP 环境的强化学习, 该算法能够保证收敛, 速度很快但不稳定。为提高强化学习的逼近函数的收敛稳定性, 对上述基本梯度算法进行了改进, 式(5)体现了权重变化 Δw , 此时的改进算法 Q学习速度较慢但非常稳定收敛。式(6)增加了权重参数 ϕ , 解决了 Q学习速度慢且能稳定收敛到优化策略的问题。

$$\Delta w = \beta(r_{t+1} + \gamma \max Q(s_{t+1}, a) - Q(s_t, a_t)) \times \left[\frac{\partial Q(s_t, a_t)}{\partial w} - \gamma \frac{\partial (\max (Q(s_{t+1}, a)))}{\partial w} \right] \quad (5)$$

$$\Delta w = \beta(r_{t+1} + \gamma \max Q(s_{t+1}, a) - Q(s_t, a_t)) \times \left[\frac{\partial Q(s_t, a_t)}{\partial w} - \phi \gamma \frac{\partial (\max Q(s_{t+1}, a))}{\partial w} \right] \quad (6)$$

2 学习分类器相关技术 (Accuracy-based Learning Classifier System, XCS)^[7-10]

XCS 由条件、动作和预测回报 3 个主要的参数组成: 条件指明 XCS 可以匹配的环境输入状态; 动作表示在该 XCS 匹配的环境状态下 XCS 的建议动作; 预测回报 P 表示该 XCS 被匹配且执行了该 XCS 所建议的动作后系统预期从环境获得的回报。预测误差 ϵ 表示预测回报 P 的误差; 适应强度 F 表示预测回报 P 的精度。

2.1 执行机制

开始分类器集合是随机生成的, 每个分类器表示为状态、

动作对。根据环境输入, 产生一个匹配环境输入的匹配规则集, 然后依据每个分类器的回报预测和适应强度值, 生成整个系统的预测列表和动作集, 最后依据分类器的动作概率选择一个作用于环境的动作。

$$P(a_i) = \frac{\sum_{cl_k \in |M|} a_i P_k \times F_k}{\sum_{cl_k \in |M|} a_i F_k} \quad (7)$$

2.2 强化机制

分类器在执行机制过程中, 每选择一个动作 a 作用于环境, 分类器从环境获得一个回报 R , 回报预测、分类器适应值强度更新如下:

$$P \leftarrow R + \gamma \max_a P(a) \quad (8)$$

式中, $0 \leq \gamma \leq 1$ 是学习效率。

$$P_j \leftarrow P_j + \beta(P - P_j) \quad (9)$$

$$\epsilon_j \leftarrow \epsilon_j + \beta(|P - \epsilon_j| - \epsilon_j) \quad (10)$$

$$k_j = \begin{cases} 1, & \text{if } \epsilon_j \leq \epsilon_0 \\ \alpha(\epsilon_j / \epsilon_0)^{-\nu}, & \text{otherwise} \end{cases} \quad (11)$$

$$k_j' = \frac{(k_j \times \text{num}_j)}{\sum_{cl_k \in |A|_{-1}} (k_k \times \text{num}_k)} \quad (12)$$

式中, 参数 ϵ_0 ($\epsilon_0 > 0$) 控制预测误差 ϵ 的冗余度; 参数 α ($0 < \alpha < 1$) 和 ν ($\nu > 0$) 是常量, 控制当 ϵ_0 被超过时精确度 k 的下降速度。动作集中的绝对精确度值 k 被转换为相对精确度值 k' , XCS 的适应度值 F 依据相对精确度值进行更新:

$$F_j \leftarrow F_j + \beta(k_j' - F_j) \quad (13)$$

2.3 规则发现机制

当匹配规则集中没有分类器与环境输入信息相匹配时, 启动规则发现机制的覆盖算法, 随机产生一个分类器。覆盖算法产生的新规则的条件部分与环境信息匹配, 该规则淘汰规则集中强度最小的规则, 存入规则集中。

3 基于梯度下降 Q学习的 XCS 设计 (XCSG)

3.1 $Q(s_{t+1}, a_{t+1})$ 到 XCS 的映射

为提高 XCS 的学习能力, 我们将梯度下降算法加入 XCS 的分类器预测更新, 建立 $Q(s_{t+1}, a_{t+1})$ 和 XCS 的映射关系。 $Q(s_{t+1}, a_{t+1})$ 和系统预测 $P(a_{t+1})$ 的关系如下:

$$Q(s_{t+1}, a_{t+1}) = P(a_{t+1}) = \frac{\sum_{cl_j \in [A]_{-1}} P_j \times F_j}{\sum_{cl_j \in [A]_{-1}} F_j} \quad (14)$$

式中, P_j 和 F_j 分别是分类器 cl_j 的预测值和适应度值。

3.2 Δw 到 XCS 的映射

在一般的函数逼近方法中, 权重矩阵 W 的权值 w 通常结合函数参数产生逼近值。而在 XCS 中, 分类器预测为函数逼近方法的关键因素。因此在 XCSG 中, 我们依据分类器 cl_j 的预测值 P_j 来计算偏导数 $Q(s_{t+1}, a_{t+1})$, 从而估计分类器 $cl_j \in |A|_{-1}$ 的梯度组件。

$$\begin{aligned} \frac{\partial Q(s_{t+1}, a_{t+1})}{\partial w} &= \frac{\partial}{\partial P_k} \left[\frac{\sum_{cl_j \in [A]_{-1}} P_j \times F_j}{\sum_{cl_j \in [A]_{-1}} F_j} \right] \\ &= \frac{1}{\sum_{cl_j \in [A]_{-1}} F_j} \frac{\partial}{\partial P_k} \left[\sum_{a_j \in [A]_{-1}} P_j \times F_j \right] \\ &= \frac{F_j}{\sum_{cl_j \in [A]_{-1}} F_j} \end{aligned} \quad (15)$$

3.3 梯度下降的强化机制

XCSG 算法的强化机制与原始 XCS 算法相比, 仅在分类器预测更新方面进行了改进。XCSG 其它参数的更新方法与 XCS 相同。当 $|A|_{-1}$ 中分类器参数更新时, 相应分类器的适

应度累计和 $F_{[A]_{-1}}$ 调整如下:

$$F_{[A]_{-1}} = \sum_{a_j \in [A]_{-1}} F_j \quad (16)$$

对每个分类器 $c_k \in [A]_{-1}$, 其预测 P_k 更新为:

$$P_k \leftarrow P_k + \beta(r + \gamma \max P(a) - P_k) \frac{F_k}{F_{[A]_{-1}}} \quad (17)$$

3.4 机器人编码设计

机器人学习环境由若干栅格组成, 栅格如已被障碍物占用, 用“0”标记表示, 目标位置用“G”表示, 空闲栅格用“ ”表示。机器人可以停在任一空闲栅格, 也可以移动到与其相连的任一空闲栅格。机器人内有 8 个声纳传感器, 分别探测与其相连的 8 个栅格的状况。传感器信息采用两位编码: 障碍物编码: 10, 目标位置编码: 11, 空闲栅格位置编码: 00, 故每个机器人的传感输入用 16 位二进制表示。每个机器人共有 8 个动作, 分别对应于相连的 8 个栅格, 假定机器人目标位于图 1 的中心位置 G。

S_0	S_1	S_2	S_3	S_4
S_5	S_6	S_7	S_8	S_9
0	0	G	S_{10}	S_{11}
0	0	0	S_{12}	S_{13}
0	0	0	S_{14}	S_{15}

图 1 机器人状态关系图

4 实验和仿真

实验环境由若干栅格组成。仿真目标是每个机器人采用 XCSG 算法随机从任一空闲栅格到达目标位置 G。我们把每个机器人所处的环境均看成有栅格组成的空间, 机器人各自的起始位置是随机的, 目标位置 G 是统一的, 当机器人到达目标栅格位置 G 时, 整个仿真过程结束, 机器人获得固定奖励 R。本文通过不同的实验环境来检验 XCSG 在多机器人强化学习中的效果。机器人数量在较大的学习空间中可任意设定, 仿真实验数据采用的是多机器人的平均数据。

实验场景 1 25 个栅格的环境地图

本实验在 25 个栅格矩形内进行(见图 1), 其中有 8 个栅格障碍物 0、一个目标栅格 G、16 个空闲栅格。分类器个体总数 $N=1500$, 从图 2 的仿真曲线看, XCS 和 XCSG 到达目标的步数和优化过程几乎是一样的, 只是在实验的开始阶段, XCS 的学习速度比 XCSG 要快。实验场景 1 的学习空间较小, 二者均能较易达到优化, 能达到非常理想的强化学习效果。

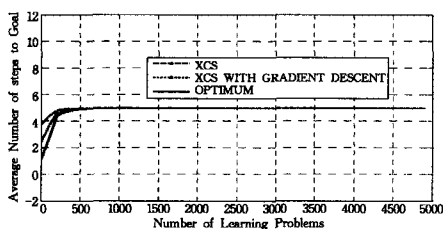


图 2 小环境下的机器人强化学习 ($N=1500$)

实验场景 2 “5”字迷宫

本实验环境为数字 5 组成的“5”字迷宫(见图 3), 从图 4 的仿真曲线上看, 在分类器个体总数 $N=2500$ 时, XCS 和 XCSG 均能实现优化操作。但在分类器个体总数 $N=3000$ 时, 图 5 仿真曲线中因 XCS 的强化学习仅能确保在分类器个体总数较少时的策略优化, 而在分类器个体总数较多时, 所有学习策略均未能达到优化, 基于梯度策略的 XCSG 强化学习方法仍然能够得到稳定的学习优化策略, 收敛速度很快, 获得了理想的强化学习效果。

0	0	0	0	0	0	0	0
0						G	0
0			0		0	0	0
0	0						0
0				0	0		0
0	0	0				0	0
0						0	0
0	0	0	0	0	0	0	0

图 3 “5”字迷宫

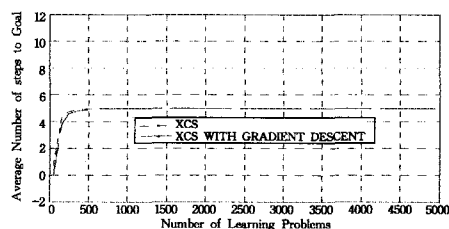


图 4 “5”字迷宫中的强化学习 ($N=2500$)

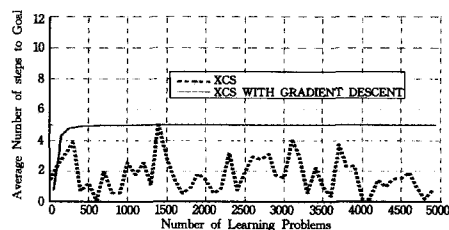


图 5 “5”字迷宫中的强化学习 ($N=3000$)

实验场景 3 “6”字迷宫

本实验环境为“6”字迷宫(见图 6), 与“5”字迷宫相比, 目标位置更加隐蔽, 这就导致机器人更多地随机行走。从图 7 看, XCS 根本不能达到优化, 而 XCSG 能够稳定快速地达到学习优化。与试验场景 2 相比, 试验场景 3 的效果更加优化, 而且能够使所有分类器个体快速实现全局优化。

0	0	0	0	0	0	0	0
0						0	G
0			0		0	0	0
0	0						0
0				0	0		0
0	0	0				0	0
0						0	0
0	0	0	0	0	0	0	0

图 6 “6”字迷宫

每个关键点的 PCA-SIFT 描述符维度是 36。标准方式 AF-SIFT1 使用 128 维描述符,而 AF-SIFT2 使用 8 方向束和 2×2 数组,其维度是 32。图 6 比较这 4 种算法的检索准确性。我们可以看到 PCA SIFT 和 AF-SIFT1 有相近的准确性,而 AF-SIFT2 比 SIFT 算法快很多。说明本文提供了一种比标准 SIFT 算法更有效的算法且其非常适合图像检索。

结束语 本文研究了一种基于综合视觉注意分析的局部特征选择的新方法,即在一幅图像中提取局部特征,按其综合显著性权值划分等级,并且仅保留最独特的局部特征。使用这种方法,在线检索速度将明显增加且精度损失很小。本文 AF-SIFT 算法对 SIFT 算法进行了改进,在图像检索的正确率和查全率方面具有很好的性能,且简单高效,适合大规模图像检索。

在以后的工作中,还要进一步研究融合多种局部特征的表达,综合多种局部特征检测算子和描述因子,建立多通道的视觉物体表示机制,以发挥不同局部特征的互补优势;还要进一步考虑图像的语义概念信息,避免造成底层局部特征和高层语义概念的脱节;同时还要进一步研究局部特征的同现性规律并建立更加有效的视觉显著性表达模型,以便进行更高效的目标检测、可靠的遥感图像分类、鲁棒的图像景观分析,提高图像检索性能。

(上接第 251 页)

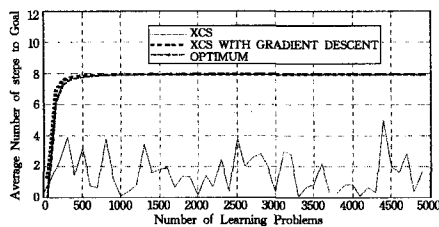


图7 “6”字迷宫中机器人强化学习($N=2500$)

结束语 本文提出了用 XCSG 来解决多机器人的强化学习问题。XCS 建立低维的状态-动作对逼近函数 Q 和状态-动作对到环境回报值的映射关系。梯度下降技术利用在线知识建立稳定的逼近函数 Q ,使状态-动作对到环境回报值的映射 Q 表格一直保持在低维状态。算法的分析和仿真表明,XCS 和梯度下降法相结合用于多机器人的强化学习是高效、稳定收敛的。

参考文献

- [1] 邵杰,杨静宇.基于多 LCS 和人工势场法的机器人行为控制[J].计算机科学,2011,38(1):264-267
- [2] 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
- [3] Wiering M. Multi-agent reinforcement learning for traffic light control[C]//Proc. 17th Int. Conf. Mach. Learn. (ICML-00). Stanford Univ. Stanford, CA, 2009:1151-1158

- [1] Ma Y F, Zhang H J. Contrast-based Image Attention Analysis by Using Fuzzy Growing[C]//ACM MM. 2003:374-381
- [2] Itti L, et al. Visual Attention and Target Detection in Cluttered Natural Scenes[J]. Optical Engineering, 2001, 40(9):1784-1793
- [3] Ke Y, et al. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors[C]//CVPR. 2004:506-513
- [4] Lowe D. Distinctive Image Features from Scale-Invariant Key-points[J]. Int'l J. Computer Vision, 2004, 2(60):91-110
- [5] Tuytelaars T. Matching Widely Separated Views Based on Affine Invariant Regions[J]. Int'l J. Computer Vision, 2004, 1(59):61-85
- [6] Mikolajczyk K, Schmid C. Performance Evaluation of Local Descriptors[J]. IEEE Trans. of PAMI, 2005, 27(10):1615-1630
- [7] Hakim A. CSIFT: A SIFT Descriptor with Color Invariant Characteristics[C]//CVPR. 2006:1978-1983
- [8] Hou X. Saliency detection: A spectral residual approach[C]//CVPR. 2007:1-8
- [9] Cheng M M, et al. Global Contrast based Salient Region Detection [C]//CVPR. 2011:456-463
- [10] 陈硕,吴成东,陈东岳.基于视觉显著性特征的快速场景配准方法[J].中国图象图形学报,2011(7):1241-1247
- [11] 程仁贵,刘书妍.基于边缘检测的影像多线自动测量算法[J].重庆理工大学学报:自然科学版,2013,27(2):89-92

- [4] Dixon P W, Corne D W, Oates M J. A preliminary investigation of modified XCS as a generic data mining tool[C]//Lanzi P L, Stolzmann W, Wilson S W, eds. LNAI, Advances in Learning Classifier Systems, vol. 2321, Berlin, Germany: Springer-Verlag, 2002:133-150
- [5] 欧世峰,高颖,赵晓晖.基于随机梯度的变动量因子自适应白化算法[J].自动化学报,2012,38(8):1370-1374
- [6] Butz M V, Goldberg D E, Lanzi P L. Gradient descent methods in learning classifier systems: Improving XCS performance in multistep problems[J]. IEEE Trans. Evol. Comput., 2005, 9(5):452-473
- [7] Bernad E, o-Mansilla, Garrell J. Accuracy-based Learning Classifier Systems: Models, analysis and applications to classification tasks[J]. Evolutionary Computation, 2003, 11(3):209-238
- [8] Hung K-T, Liu J-S, Chang Y-Z. Smooth path planning for a mobile robot by evolutionary multiobjective optimization[C]//IEEE Int. Symposium on Computational Intelligence in Robotics and Automation. Jacksonville, Florida, June 2007
- [9] Butz M V, Lanzi P L, Wilson S W. Function approximation with XCS: Hyperellipsoidal conditions, recursive least squares, and compaction[J]. IEEE Trans. Evol. Comput., 2008, 12(3):355-376
- [10] Bagnall A J, Cawley G C. Learning classifier systems for data mining: A comparison of XCS with other classifiers for the Forest Cover dataset[C]//Proc. IEEE/INNS Int. Joint Conf. Artificial Neural Netw. vol. 3, Portland, OR, 2003:1802-1807