第 34 卷 第 11 期 计 算 机 工 程 2008 年 6 月
Vol.32 No.11 Computer Engineering June 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)11—0219—03 文献标识码：A 中图分类号：TP18

# 共享经验的多主体强化学习研究

焦殿科[1]，石 川[2]
(1. 辽宁工业大学计算机科学与工程学院，锦州 121001；2. 北京邮电大学北京市智能软件与多媒体重点实验室，北京 100088)

**摘 要：**合作多主体强化学习的关键问题在于如何提高强化学习的学习效率。在追捕问题的基础上，该文提出一种共享经验的多主体强化学习方法。通过建立合适的状态空间使猎人共享学习经验，根据追捕问题的对称性压缩状态空间。实验结果表明，共享状态空间能够加快多主体强化学习的过程，状态空间越小，Q 学习算法收敛越快。

**关键词：**合作多主体；强化学习；Q 学习算法；状态空间

# Research on Multi-agent Reinforcement Learning with Sharing Experience

JIAO Dian-ke[1], SHI Chuan[2]
(1. College of Computer Science & Engineering, Liaoning University of Technology, Jinzhou 121001;
2. Beijing Key Laboratory of Knowledgeware and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100088)

**【Abstract】** How to improve the efficiency of reinforcement learning is the key problem of reinforcement learning with multi-agent collaboration. This paper proposes a method of multi-agent reinforcement learning with sharing experience based on the research to pursuit problem. By applying this method the hunters can share the learning experience through constructing the appropriate state space. It further compresses the state space according to the symmetry character of pursuit problem. Experimental results show that sharing state space can expedite the process of multi-agent reinforcement learning. The smaller the state space is, the faster Q learning algorithm convergence will be.

**【Key words】** multi-agent collaboration; reinforcement learning; Q learning algorithm; state space

强化学习方法是一种无导师机器学习方法，采用不断“试错-修改”过程，不需要完备的推理过程，只需要通过反馈信息，以合适的算法强化好的行为，弱化差的行为，最终收敛到最优行为[1]。多agent强化学习是强化学习研究中的一项。在多agent系统中，环境在多个agent的联合动作下进行状态的迁移。对于单个agent来讲，由于其只能确定自身agent的行为动作，因此体现出一种行为动作上的“部分感知”，从而产生非标准马尔可夫环境。多agent强化学习机制被广泛应用到各个领域，例如游戏、邮件路由选择、电梯群控系统以及机器人设计等。针对合作多主体强化学习问题，本文分析主体共享状态空间的方式对Q学习性能的影响。对于猎人围捕问题，提出了 3 种状态空间共享方式，分析状态空间对算法性能的影响。

## 1 相关工作

多主体系统的学习不是单主体学习的简单增强。事实上，多主体的学习过程是相当复杂的，直接依赖于多个主体的存在和交互。Q 学习是一种与模型无关的强化学习方法，它不仅应用于单个主体的强化学习中，而且广泛应用于多主体强化学习中。

### 1.1 猎人追逐问题

猎人围捕问题是一个经典的人工智能问题，由于该问题具备多主体系统的多种特性，对于现实世界的问题具有通用性，而且易于扩充，因此经常被用来研究多主体系统的学习、协作行为、通信等问题[2]。首先假定网格世界中存在 4 个猎人和 1 个猎物，其中 4 个追逐猎人被视为 1 个协作团队，猎人分布在 4 个角上，而猎物在网格世界的中央。猎物在网格世界中走动，其模型不为猎人所知，猎物被视为猎人团队所处的动态环境的一部分。4 个猎人随之走动，每个猎人都是一个决策单元，同时也是一个学习单元。

很显然，团队目标的实现必须要求猎人之间进行有效的协作，该协作团队试图通过协同强化学习，寻求最优的追捕猎物的联合行为策略。在这个问题中只要有一个猎人追逐到猎物，则完成了追捕任务。

### 1.2 合作多主体强化学习的进展

猎人围捕问题属于合作多主体强化学习。合作多主体强化学习的基本思想在于“在主体选择动作之前，相互交互，产生更新以后的值函数，而动作的选择基于新的值函数”。早在 20 世纪 90 年代初，文献[3]指出合作多主体强化学习中相互交互(交换信息)是最有效的方法之一，并给出了 3 种主要的实现方法：

(1)交换每个 agent 感知的信息状态；
(2)交换 agent 学习的经验片段；
(3)交换学习过程中的策略和参数等。

**作者简介：**焦殿科(1953－)，男，副教授，主研方向：强化学习，进化计算，计算机网络；石 川，博士
**收稿日期：**2008-04-21 **E-mail：**jiaodianke@163.com

2004 年，文献[4]给出交换建言方法。与单 agent 学习相比，以上方法可以有效地提高学习速度。文献[5]采用新的状态行为的知识表示方法使状态行为空间得到缩减，采用相似变换和经验元组的共享似学习效率得到了提高。文献[6]提出了基于 Agent 团队的强化学习模型。该模型引入主导 agent 的角色作为团队学习的主角，并通过 agent 角色的变幻实现整个团队的学习。

## 2 共享经验的多主体强化学习

在猎人追捕问题中，每个猎人的局部状态是指猎人与猎物对应的目标位置的相对位置，局部行为是指猎人的走动情况。每个猎人只需要关心自己的局部状态和局部行为。环境的全局状态由各个局部状态组成。当环境处于某个状态时，猎人采取动作之后的状态转移和瞬时回报与猎人的历史状态和行为无关，只与当前的状态和行为相关，也就是说，猎人采取的行动与猎人的历史状态无关，只能根据当前的感知情况来行动，因此，这样的环境具有明显的马尔科夫特性。而且由于猎物处于运动状态，无法事先给出每个猎人可能的状态迁移概率函数，每个状态下的行为的汇报函数也无法给出，对于这样的环境，利用强化学习方法解决是一个理想的选择。

在 Q 学习算法中，状态行为对的 Q 值是指导学习的线索，学习的目的就是要得到 Q 值最大的状态行为对。但是，这种学习方法需要通过重复访问状态行为空间。当行为状态空间较大时，收敛速度较慢。本文使用如下几种方法表示多主体的行为状态空间。

### 2.1 以猎人为中心建立状态空间

以猎人为坐标原点建立坐标，则猎物相对于猎人的相对坐标为 $(x_0 - x, y_0 - y)$。状态定为猎物相对于猎人的相对坐标 $(x_0 - x, y_0 - y)$。

如图 1 中的坐标所示，0 表示猎物，1~4 表示猎人，可以以猎人或猎物为中心建立状态空间。猎物在猎人 1~猎人 4 所在的不同状态空间中的状态分别是(-3,-2), (2,-3), (3,2)和(-2,3)。每个猎人都有自己的坐标系。同一个猎物在不同坐标系中位置不同。猎人之间无法共享经验。有多少猎人就有多少状态空间。每个状态空间的大小为网格的大小。

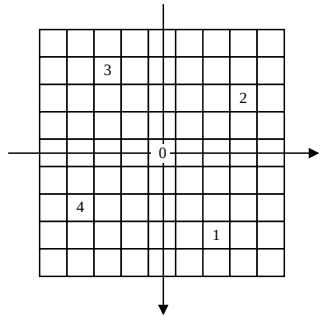



**图 1 根据对称性建立的坐标**

### 2.2 以猎物为中心建立状态空间

以猎物为坐标原点建立坐标，则猎人相对于猎物的相对坐标为 $(x - x_0, y - y_0)$。状态定为猎人相对于猎物的相对坐标 $(x - x_0, y - y_0)$。如图 1 所示，猎人 1~猎人 4 相对于猎物的状态空间中的状态是(3,2),(-2,3),(-3,-2)和(2,-3)。每个猎人可以相对于猎物有一个状态空间，猎人之间不共享状态空间。这种建立状态空间的方法和以猎人为中心的方法没有多大区别，有多少个猎人就有多少状态空间。

### 2.3 根据对称性压缩状态空间

以猎物为中心建立状态空间，所有猎人都可以使用这个统一的坐标，不同猎人可以共享状态空间。坐标将网格划分成了 4 个象限，其中的状态有某种对称性。在图 1 中，1 旋转 90º到 2，旋转 180º到 3，旋转 270º到 4。同样他们的动作也具有这种旋转对称性。例如：1 的向左运动，分别对应于 2、3、4 的向下、向右、向上运动。猎人 1~猎人 4 在以猎物为中心的状态空间中的状态分别为 $(x_1, y_1)$, $(x_2, y_2)$, $(x_3, y_3)$, $(x_4, y_4)$，他们分别位于第 1 象限~第 4 象限。

如果将他们压缩到第 1 象限内的状态空间，其他 3 个象限的状态与第 1 象限的状态对应关系为

$$\begin{cases} x_2 = -y_1 \\ y_2 = x_1 \end{cases} \quad \begin{cases} x_3 = -x_1 \\ y_3 = -y_1 \end{cases} \quad \begin{cases} x_4 = y_1 \\ y_4 = -x_1 \end{cases}$$

同样 4 个象限中的动作也有类似的关系，对应关系如下：

(1)第 1 象限：上、下、左、右。

(2)第 2 象限：右、左、上、下。

(3)第 3 象限：下、上、右、左。

(4)第 4 象限：左、右、下、上。

利用上面的方法可以将 4 个象限的状态统一到第 1 象限，这样进一步地减少了 3/4 的状态空间。以图 1 为例，4 个猎人追捕一个猎物，状态变化关系见表 1。

**表 1 猎人坐标的转换关系示例**

| 猎人 | 绝对坐标 | 对猎物的相对坐标 | 所处象限 | 压缩坐标 |
|---|---|---|---|---|
| 1 | (7,6) | (3,2) | 1 | (3,2) |
| 2 | (2,7) | (-2,3) | 2 | (3,2) |
| 3 | (1,2) | (-3,-2) | 3 | (3,2) |
| 4 | (6,1) | (2,-3) | 4 | (3,2) |

## 3 实验研究

为了验证不同状态空间的表示方法对合作效率的影响，本文研究 30×30 大小的猎人围捕实验。还是依照猎物在中央，猎人在四周的模型来建立。实验测试两种情况：4 个或 8 个猎人追捕猎物。当有 4 个猎人时，初始位置位于 4 个角上；当 8 个猎人时，初始位置平均的分布在边缘地区。猎物位于中央位置。猎人可以在网格中全范围地感知猎物。猎人和猎物可以在网格环境中上下左右随意移动。学习任务是：猎人要捕获到猎物。如果主体朝边界移动，它将留在原地，并没有惩罚。折扣因子为 $\gamma = 0.9$ 并且没有立即回报。只有猎人捕获到猎物才有回报，回报值为 5。算法中学习率为 $\alpha = 0.1$，并且使用 $\varepsilon$-贪婪策略，$\varepsilon = 0.1$。

实验采用了 3 种不同的方式建立状态空间：

(1)猎人以猎物为中心建立自己的相对坐标作为状态空间(IndS)。

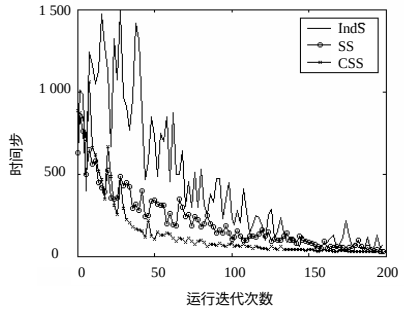
(2)以猎物为中心，所有猎人建立统一共享的相对坐标作为状态空间(SS)。

(3)在共享相对坐标的基础上，利用对称性将状态空间压缩到第 1 象限(CSS)。

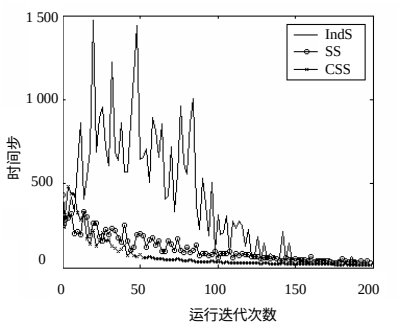
这 3 种算法采用相同 Q 学习算法，算法独立运行 10 次，结果是多次运行的平均值。实验记录每次猎人抓到猎物的运行时间步。在每个时间步，猎物和每个猎人都移动一步。

图 2 反映了 3 种状态空间情况下，猎人捕获到猎物所需的时间步。其中，图 2(a)显示 4 个猎人捕获问题，图 2(b)显示 8 个猎人捕获问题。通过实验可以发现，采用相对坐标作

为状态空间，3 种方法使用 Q 学习算法都是收敛的。采用共享空间的方法，比不采用共享空间的方法要快得多，如果采用压缩空间的方法，收敛速度更快。




(a)4 个猎人的捕获结果




(b)8 个猎人的捕获结果

**图 2　3 种不同的状态空间建立方法对算法的影响**

图 3 反映了采用 CSS 共享状态空间、不同猎人数目捕获猎物所需的时间步。实验结果也验证：尽管猎物随机移动，但是猎人越多，就能越快捕获到猎物。因为所有猎人都共享状态空间，状态空间记录的学习路径数随猎人数的增加呈线性增长的趋势。学习路径越多，Q 学习可以越快收敛到最优解。

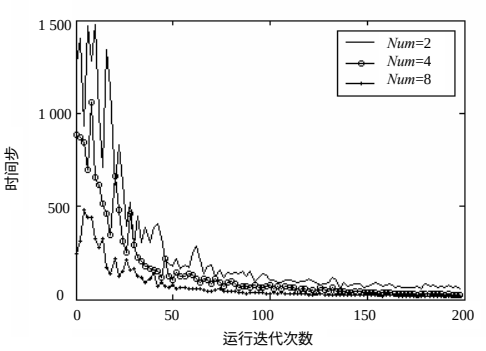



**图 3　不同猎人数目捕获猎物的时间步比较**

## 4　结束语

合作多主体强化学习系统强调如何利用分布式强化学习来提高强化学习速度。通过猎人围捕问题，本文深入研究了共享经验对多主体强化学习的学习速度的影响。通过实验发现，共享的状态空间加快了学习的速度；空间状态越小，学习速度也越快。

### 参考文献


[1] Mitchell T M. 机器学习[M]. 曾华军, 张银奎, 译. 北京: 机械工业出版社, 2003.

[2] Nitschke G. Emergence of Cooperation in a Pursuit-evasion Game[C]//Proc. of the 18th International Joint Conference on Artificial Intelligence. Acapulco, Mexico: [s. n.], 2003: 639-646.

[3] Tan M. Multi-agent Reinforcement Learning: Independent vs Cooperative Agents[C]//Proc. of the 10th International Conference on Machine Learning. Amherst, MA: [s. n.], 1993: 330-337.

[4] Nunes L. Cooperative Learning Using Advice Exchange[M]. Berlin, Heidelberg, Germany: Springer-Verlag, 2003: 33-48.

[5] 王长缨, 尹晓虎, 鲍翊平. 一种共享经验元组的多 agent 协同强化学习算法[J]. 模式识别与人工智能, 2005, 18(2): 234-239.

[6] 蔡庆生, 张　波. 一种基于 Agent 团队的强化学习模型与应用研究[J]. 计算机研究与发展, 2000, 37(9): 1086-1093


（上接第 218 页）

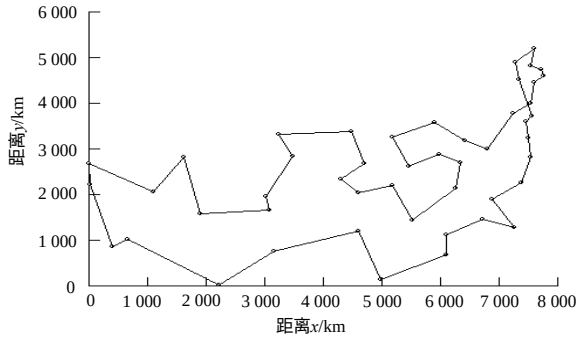



**图 2　基于模拟退火的粒子群算法解 att48 的最优解**

## 5　结束语

本文用模拟退火算法和粒子群算法相结合的算法思想解决典型的离散优化问题。本算法的最大特点是算法简单、实现容易且效果优于一般的遗传、蚁群等算法，能使粒子群算法更好地应用于离散领域，具有较高的实用价值。

### 参考文献


[1] Eberhart R C, Kennedy J. A New Optimizer Using Particles Swarm Theory[C]//Proc. of the 6th Int'l Symposium on Micro Machine and Human Science. Nagoya, Japan: [s. n.], 1995.

[2] Shi Yuhui, Eberhart R C, Fuzzy Adaptive Particle Swarm Optimization[C]//Proc. of Congress on Evolutionary Computation. San Francisco, USA: IEEE Press, 2001.

[3] Lovbjerg M, Rasmussen T K, Krink T. Hybrid Particle Swarm Optimizer with Breeding and Subpopulation[C]//Proceedings of Evolutionary Computation Conference. San Francisco, USA: [s. n.], 2001.

[4] Ciuprina G, Ioan D, Munteanu I. Use of Intelligent Particle Swarm Optimization in Electmmagnetics[J]. IEEE Trans. on Magnetics, 2002, 38(2): 1037-1040.

[5] 康立山, 谢　云, 尤矢勇, 等. 模拟退火算法[M]. 北京: 科学出版社, 1994: 50-97.

[6] Clerc M. Discrete Particle Swarm Optimization Illustrated by the Traveling Salesman Problem[Z]. [2007-04-20]. http://www.mauriceclerc.net.

[7] 高　尚, 韩　斌, 吴小俊, 等. 求解旅行商问题的混合粒子群优化算法[J]. 控制与决策, 2004, 19(11): 1286-1289.