学习过程中共享经验的Q学习算法的研究

乔 林 罗 杰
(南京邮电大学自动化学院 南京 210046)

摘 要 主要以提高多智能体系统中Q学习算法的学习效率为研究目标,以追捕问题为研究平台,提出了一种基于共享经验的Q学习算法。该算法模拟人类的团队学习行为,各个智能体拥有共同的最终目标,即围捕猎物,同时每个智能体通过协商获得自己的阶段目标。在学习过程中把学习分为阶段性学习,每学习一个阶段,就进行一次阶段性总结,分享彼此好的学习经验,以便于下一阶段的学习。这样以学习快的、好的带动慢的、差的,进而提升总体的学习性能。仿真实验证明,在学习过程中共享经验的Q学习算法能够提高学习系统的性能,高效地收敛于最优策略。

关键词 Q学习算法,MAS,围捕问题,共享经验
中图法分类号 TP181 **文献标识码** A

Research on Q Learning Algorithm with Sharing Experience in Learning Process

QIAO Lin LUO Jie
(College of Automation, Nanjing University of Posts & Telecommunications, Nanjing 210046, China)

Abstract The aim of the research is to improve the efficience of multi-agent Q-learing algorithm. This paper proposed a method of multi-agent Q-learning with sharing experience based on the pursuit problem. This algorithm simulats human behavior of a learning team, and all agents share a common utimate goal of capturing the prey, at the same time every agent gets their own milestones through negotiations. The learning process is divided into some stages. After a learning stage, there will be a stage summary. Then good learning experience will be shared with each other in order to facilitate the next stage of learning. The agents who learn fast and well can help the ones who learn slow and not well, so in this way the performance of the system is enhanced. The simulation results prove that the Q-learning algorithm with sharing experience in learning process can improve the performance of learning systems and efficient convergence to the optimal strategy.

Keywords Q-learning algorithm, MAS, Pursuit problem, Sharing experience

1 引言

增强学习又称为强化学习或再励学习,是近年来机器学习和智能控制领域的前沿和热点,与监督学习和无监督学习并列为3大类机器学习方法。增强学习强调以不确定条件下序贯决策的优化为目标,是复杂系统自适应优化控制的一类重要方法[1]。增强学习是一种试错学习,Agent通过与环境不断交互,来发现由环境状态到动作的一种映射关系。在这一过程中根据执行动作得到的结果来给予一定的奖惩,通过不断地累积奖惩来激励智能体选择最优动作。目前,增强学习主要有Sutton首次提出的TD(λ)算法[2]、Watkins首次提出的Q学习算法[3]以及Rummery提出的Sarsa算法[4]。其中发展最快、应用最为普遍的是Q学习算法。

基于共享经验的Q学习算法的研究,一些学者已经做了一定的工作。王长缨、尹晓虎提出了一种共享经验元组的多Agent协同强化学习方法[5],即多个智能体共享一张表,彼此互斥地更新这张表,以求达到经验共享。焦殿科、石川提出了共享经验的多主体强化学习方法[6],即通过分析状态空间的对称性来压缩状态空间,从而实现部分经验共享。本文提出的经验共享是对学习过程中各个智能体所学到的经验知识阶段性地进行共享,每个智能体都有自己独立的状态空间,真正实现在学习过程中共享经验。

本文以围捕问题为平台研究多智能体系统中的Q学习算法。4名猎人通过连续地学习与合作来围捕一个不断逃跑的猎物。在这一过程中,通过模拟最具智能性的人类思维来完成学习与合作。起初,猎人对环境一无所知,他们有目标而无方向地在区域中游走,以获取相应的环境知识。接着,所有猎人有目标且有方向性地开始学习。在已有环境的基础上,猎人通过不断地制定目标,学习、分享经验,再学习,再分享经验的循环过程来实现最终的目标,即围住猎物。本文提出的基于经验共享的Q学习算法正是以提高算法的效率为目标而进行的改进工作,具体体现为成功围捕猎物次数的增加和所需步数的减少。同时,对奖惩方法进行了改进,根据具体需要开发了多角度评分方法,该方法避免了采用联合动作和联

到稿日期:2011-06-07 返修日期:2011-09-30
乔 林(1986—),女,硕士,主要研究领域为智能机器人、模式识别与人工智能等,E-mail:huhu0714@yahoo.com.cn;罗 杰(1963—),男,博士,教授,主要研究领域为分布式智能控制、群体智能。

合奖惩的奖惩方法，从而缓解了维数灾难，提高了搜索效率。

2 基于共享经验的Q学习算法

2.1 Q学习算法

Q学习算法是增强学习的主要算法之一，是一种与模型无关的学习算法。Q学习算法是基于马尔科夫决策过程(Markov Decision Process，MDP)的递增式动态规划算法。马尔科夫决策过程由五元组组成，分别是[7]：

S：环境有限状态集合

A：有限动作空间集合

T：状态转移函数

R：回报函数

V：评价函数

在马尔科夫决策过程中，智能体每进行一步操作都需要先观察当前所处的状态 s_t，$s_t \in S$，然后从动作集合 A 中以某种控制策略选择动作 a_t 执行，进而转移到下一个状态 s_{t+1}，从而得到一个即时回报 r_t。在Q学习算法中，每个 $Q(s_t,a_t)$ 都对应一个 Q 值，这个 Q 值就是按照所选择的策略持续执行而得到的回报 r 的总和。我们所选择的策略就是要让这个累积回报最大。Q学习算法中，累计回报定义如下[8]：

$$Q(s_t,a_t)=Q(s_t,a_t)+\alpha[r+\gamma\max Q(s_{t+1},a_{t+1})-Q(s_t,a_t)] \tag{1}$$

式中，α 为学习率($\alpha>0$)，γ 为折扣因子($0\leqslant\gamma<1$)。算法中最优策略为 $\pi^*=\max Q(s_{t+1},a_{t+1})$，即选择具有最大回报值的动作。在本文中以一定的概率 $\varepsilon(0<\varepsilon<1)$ 选择具有最大回报值的动作，以 $(1-\varepsilon)$ 的概率随机选择动作，使智能体尽可能覆盖所有可能的状态。这样可以减少算法收敛于次优解而非最优解的情况。

2.2 在过程中共享经验的Q学习算法

多智能体系统中的Q学习算法同单体的Q学习算法相似，只是智能体的状态变成了联合状态 $\vec{s}=(s^1,s^2,\cdots,s^n)$，动作变成了联合动作 $\vec{a}=(a^1,a^2,\cdots,a^n)$，回报则变成由联合状态和联合动作而得到的回报。这样，每次多智能体选择动作时，都需要搜索联合状态和联合动作。当这个学习系统很大时，从状态到动作的映射集合将会随着智能体的增加呈指数增长，这就是维数灾难问题。

在本文中，各个智能体的学习是部分独立的，即动作的选择、执行以及得到回报均是独立的。在执行任务之初，所有智能体会进行一次协商来分配各自的任务。在环境发生变化时，所有智能体会进行再次协商，重新分配任务，直到共同的目标实现为止。智能体在学习一个阶段后会彼此分享一下经验，共享各自在前一阶段学习到的好的成果，以便于在下一阶段中更好地学习。其定义式为：

$$Q(s_t^i,a_t^i)=\max(Q(s_t^1,a_t^1),Q(s_t^2,a_t^2),\cdots,Q(s_t^n,a_t^n))$$
$$i=[1,2,\cdots,n] \tag{2}$$

即各个智能体在状态 s_t 下执行动作 a_t 得到累积回报值，从这些回报值中选出最大值，分享给其他智能体。通过在多智能体系统学习的过程中共享经验，可以让学习快的智能体带动学习慢的智能体，整个系统的学习效率也将有所提高。

智能体开始学习前，对环境一无所知。为了后续更好地学习，在学习之初，智能体彼此之间仍先制定目标，但不选择具有最大回报值的动作，而是随机地游走，尽可能多地覆盖各种状态，并得到一定的回报。这些环境知识减少了智能体因为对环境的不了解而走弯路，最终收敛于次优解的情形。

针对本文的问题，提出了双奖惩标准的奖惩方法。从不同角度给予奖惩，根据具体需要选择奖惩方法，也就是说将有两个状态集。这样的优点在于可以通过判断减少智能体的搜索范围，提高搜索效率，缓解因为联合动作和联合状态而带来的维数灾难问题。具体算法如下：

第一步　初始化状态集 S、状态集 S'、动作集 A，以及 $Q(s,a)$。

第二步　在学习初期重复执行以下操作：

随机选择动作 a 并执行；

获得立即回报 r；

更新 $Q(s,a)$。

第三步　重复执行以下操作：

观察当前状态 s；

根据策略选择动作 a 并执行；

获得立即回报 r；

按策略更新 $Q(s,a)$；

判断是否满足终止条件。

第四步　共享经验，即共享 $Q(s,a)$，继续执行第三步操作。

3 基于共享经验的围捕问题

3.1 围捕问题模型

猎人围捕问题是一个经典的人工智能问题。该问题由于具备多智能体系统的多种特性，而且易于扩充，因此经常被研究人员用来研究多智能体系统的学习、协作、通信、竞争性协进化等多种问题[5]。在本文中，背景环境是一个 7×7 的无边界的格栅世界，猎人4位，猎物1个。

猎物被分配在中心位置，猎人在其他位置随机分配，如图1所示。其中星形表示猎人，实心点表示猎物。本文中将猎物作为原点，每一个格栅代表一种状态。猎人可选择的动作有上移、下移、左移、右移、不动这5个。在学习过程中，猎人不可以同时到达同一个位置，若试图到达同一位置，则被强行退回原位。在开始围捕前，4位猎人会根据现在所处的位置与坐标轴的夹角来选择自己的目标位置。与X正半轴夹角最小的目标为猎物东边的位置，与X负半轴夹角最小的目标为猎物西边的位置。与Y正半轴夹角最小的目标为猎物北边的位置，与Y负半轴夹角最小的目标为猎物南边的位置。只要猎人们都到达目标位置，则围捕成功。

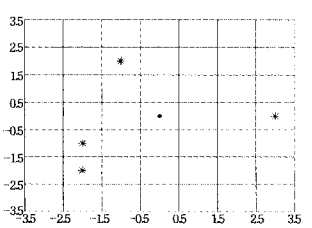

图1　位置分布图

本文中有两套奖惩办法。将整个围捕环境划分成4个扇

形区域，如图 2 所示。如果猎人所处的位置与其目标位置在同一区域，则以猎人与猎物之间的距离变化作为奖惩标准。如果猎人所处的位置与其目标位置不在同一区域内，则以猎人与目标区域夹角的变化作为奖惩标准。因此有两个状态集，一个存放由距离变化作为评判标准的奖惩值，一个存放由夹角变化作为评判标准的奖惩值。在猎人抓捕猎物的过程中，首先做的工作是对猎物形成一个包围圈，接着以距离为准绳，缩小包围圈，直至抓住猎物。

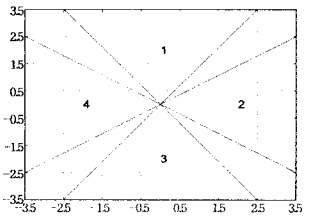

图 2　区域图

为了模拟现实场景中的围捕情境，可让猎物的反应比猎人慢一些。猎物在一个比围捕范围小的菱形区域中自由游走，猎人从大一点的范围中围捕猎物。当猎物的位置发生变化后，猎人会再次协商，重新制定目标。接着以猎物的位置作为新的原点，继续围捕。

3.2　仿真实验及结果分析

基于共享经验的 Q 学习算法的实验参数为 $\alpha=0.1$，$\gamma=0.95$，$\varepsilon=0.8$。$r=\{-1,0,1\}$，分别对应{相对猎物距离变远，相对猎物距离不变，相对猎物距离变近}，初始 Q 值均设为 0。猎人最多可以移动 100 步。4 个猎人中，有一个移动超过 100 步，则此次围捕失败。实验共进行了 100 次围捕行动，每一次实验都是在前一次实验的基础上继续学习。每一次实验开始时猎人的位置都会被随机地分配。

本实验为对比实验，第一组实验采用未改进的 Q 学习算法，第二组采用基于共享经验和双奖惩标准的 Q 学习算法。在第二组实验中，100 次围捕的前 10 次，猎人选定目标却不按策略地随机行走，后 90 步则按照改进后的算法执行。图 3 展示了第二组实验中某一次实验的围捕过程。图 4、图 5 展示了两组实验的结果。

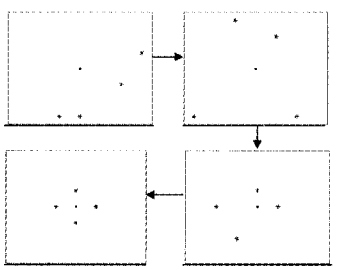

图 3　围捕过程

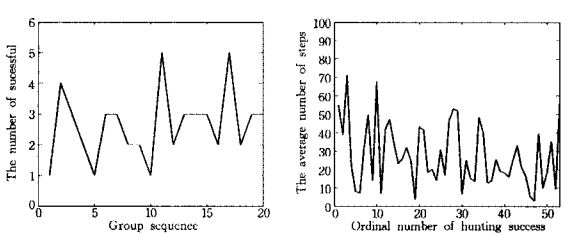

图 4　标准 Q 学习算法实验结果

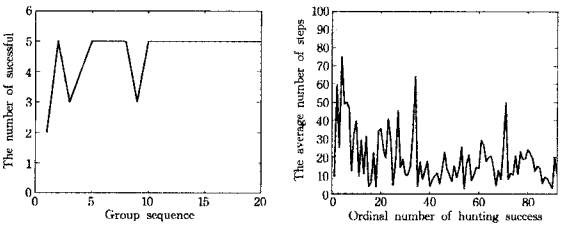

图 5　基于共享经验的 Q 学习算法实验结果

100 次实验中，每 5 次实验作为一个统计小组，记录每个统计小组中成功围捕的次数，得到 20 组实验数据。如图 4、图 5 中的左图所示，横坐标为实验的组次，纵坐标为每个统计小组中成功围捕的次数。记录围捕成功的实验中每个猎人行走的步数，并对 4 名猎人行走的步数取平均值，如图 4、图 5 的右图所示。横坐标是成功围捕的序次，纵坐标是每个猎人当次成功围捕行走的平均步数。

从图 3 可以看出，围捕过程中猎人以形成包围圈作为首要目标，以缩小包围圈为次要目标，最终成功抓住猎物。本试验中，环境改变频繁，猎物和猎人活动中存在一定的随机性。图 4 和图 5 都存在一定的波动性，因此从趋势上来比较两种算法。首先从图 4 和图 5 的左图中可以看出，两种算法都使围捕成功的次数有所增加，呈上升趋势。而从图 4 和图 5 的右图中可以看出，改进后的算法成功的总次数多于未改进的算法：改进前为 50 多次，改进后为 80 多次。两种算法都使围捕成功所需的步数有所下降，整体呈下降趋势，但改进后的算法下降趋势较未改进的更为明显。除去因为随机原因而突出的几个点外，改进后的 Q 学习算法趋于一个相对稳定的状态。

为了证明改进后算法的稳定性，我们做了大量统计对比实验。对标准 Q 学习算法和改进后的 Q 学习算法各做了 20 轮实验，每一轮为 100 次围捕过程。记录围捕成功的实验中 4 个机器人行走的步数，图 6 的横坐标为实验的轮数，纵坐标为该轮实验的 4 个猎人行走的平均步数。从图 6 中可以很明显地看到，改进后的 Q 学习算法总体趋势较未改进的平稳，所需要的平均步数也较改进的 Q 学习算法减少了 5 到 10 步。

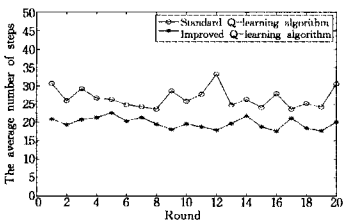

图 6　两种算法性能比较

结束语　本文提出了多智能体系统中基于经验共享的 Q 学习算法。算法模拟人类在学习过程中阶段性地进行经验总结，学得多、学得快的将自己好的经验分享给落后的学习者，所有人一起进步，因而也能较快地实现个人目标和集体目标。从多角度、多方位给予奖惩，更能减少智能体因为单个条件的约束而进行许多不必要的重复工作。从结果中可以看出，基于经验共享和双奖惩标准的 Q 学习算法在整体性能和收敛速度上都较标准 Q 学习算法有一定程度的提高。改进后的

Q学习算法能够很好地适应动态的环境，更高效地收敛于最优策略。

参 考 文 献

[1] 徐昕. 增强学习与近似动态规划[M]. 北京：科学出版社，2010

[2] Sutton S. Learing to predict by the methods of temporal difference[J]. Machine Learing，1998(3)：9-44

[3] Watkins C J C H，Dayan P. Technical note：Q-learing[J]. Machine Learing，1992，8(3/4)：279-292

[4] 张芳. 面向多移动机器人系统的再励学习方法研究[D]. 上海：上海交通大学，2002

[5] 王长缨，尹晓虎，鲍翊平，等. 一种共享经验元祖的多 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] 焦殿科，石川. 共享经验的多主体强化学习研究[J]. 计算机工程，2008，34(11)：219-221。

[7] 刘杰. 基于强化学习的多机器人围捕策略的研究[D]. 长春：东北师范大学，2009

[8] 黄炳强. 强化学习方法及其应用研究[D]. 上海：上海交通大学，2007

[9] 潘莹，李德华，梁京章，等. 基于实用推理的多智能体协作强化学习算法[J]. 华中科技大学学报，2010，38(4)：54-57

[10] Ito K，Imoto Y，Taguchi H，et al. A study of reinforcement learning with knowledge sharing[C]// Proceedings of the 2004 IEEE International Conference on Robotics and Biomimetics. Japan：Okayama University Digital Information Repository，2004：175-180

[11] 毛俊杰，刘国栋. 基于先验知识的改进强化学习及其在 MAS 中的应用[J]. 计算机工程与应用，2008，44(24)：156-158

(上接第 197 页)

也是只有一帧的运动，其面部形态如图 8 所示。

2)悲伤的时候说哦。这两个运动的合成与上一个实验类似，合成运动的面部形态如图 9 所示。

3)对于人的复杂情感，比如悲喜交加，这种情况下很难描述出它的面部形态，也就很难应用正向的运动学控制生成方法和示教再现的面部运动生成方法。这种情况下，将两个运动合成，是一种较方便生成对应人复杂情感的面部运动的方案。定义悲喜交加为高兴与悲伤的混合，其合成运动是只有一个关键帧的运动。其关键帧的面部形态如图 10 所示。

图 8 高兴地说“哦”

图 9 悲伤地说“哦”

图 10 “悲喜交加”

结束语 本文在多维面部运动的形式化描述模型中，提出了运动的控制约束条件，然后采用了正向运动学和示教再现两种生成面部运动的方法。正向运动学方法分段控制运动，较精确地生成了各种静态面部形态以及动态面部运动，减少了非自然表情出现的概率。示教再现的方法根据面部运动的轨迹点，采用逆运动学方法计算目标点对应的面部形态向量，拟合出向量运动曲线；再通过正向运动学的方法进行展现，使设计者可以不用考虑操作细节；最后，将产生的基础运动单元合成复杂面部运动生成，简化了面部运动的生成过程，提高了面部运动生成效率。实验表明，本文方法不仅有助于生成逼真的表情，还有助于弥补过程控制缺乏的问题，帮助人们生成了面部运动过程的中间细节，实现了运动的重用。

参 考 文 献

[1] Duchenne G B，Cuthbertson R A. The Mechanism of Human Facial Expression[M]// Cuthbertson R A，ed. Cambridge：Cambridge University Press，1990

[2] Gray H，Lewis W. Anatomy of the Human Body [M]. New York：Bartleby，2000

[3] Ekman P，Friesen W V，Hager J. Facial Action Coding System [M]// Alto P. CA：Consulting Psychologists，1978

[4] Brooke N M，Summerfield Q. Analysis synthesis and perception of visible articulatory movements [J]. Journal of Phonetics，1983，11(1)：63-76

[5] Summerfield Q. Roles of the Lips and Teeth in Lipreading Vowels[C]// Proceedings of the Institute of Acoustics. 1984

[6] Summerfield Q. Lipreading and Audio-Visual Speech Perception [J]. Philosophical Transactions on Biological Sciences，1992，335：71-78

[7] Breazeal C，Scassellati B. How to build robots that make friends and influence people[C]// International Conference on Intelligent Robots and System. Piscataway：IEEE Press，1999：858-863

[8] Breazeal C，Fitzpatrick P. Social amplification of animate vision [C]// Proceedings of the AAAI Fall Symposium on Society of Intelligence Agents—The Human in the Loop. AAAI Press，2000

[9] Hegel F，Muhl C，Wrede B. Understanding Social Robots[C]// Second International Conferences on Advances in Computer-Human Interactions. 2009：169-174

[10] 李旭东，张振跃. 人脸表情的形变线性拟合方法[J]. 自动化学报，2008，34(5)：593-597

[11] Lu Xiao-guang，Jain A K. Deformation Modeling for Robust 3D Face Matching[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2008，30(8)：1346-1357

[12] Caramero L，Fredslund J. I show you how I like you—can you read it in my face? [J]. IEEE Transactions on Systems，Man and Cybernetics，2001，31(5)：454-459

[13] 吕梦雅，张志刚，唐勇，等. 基于肌肉模型的眼睛动画研究[J]. 系统仿真学报，2008，20(20)：5573-5580

[14] Yang B，Jia P. A Facial Expression Model for Human-like Agent [C]// Proceeding of 2006 IEEE Conference on Robotics，Automation and Mechatronics. 2006：1-6