MAS中基于多奖惩标准的Q学习算法研究

乔 林 罗 杰
(南京邮电大学自动化学院 南京 210046)

摘 要 传统的Q学习算法是基于单奖惩标准的。基于单奖惩标准的Q学习算法往往不能适应multi-agent system(MAS)面对的复杂变化的环境与状态,相反可能还会制约学习效率。提出的基于多奖惩标准的Q学习算法能够较好地适应复杂变化的状态与环境,分阶段完成任务,不同阶段使用不同的奖惩标准,能够快速地完成阶段目标。以三维世界中的围捕问题为仿真平台,增加了围捕的难度和状态环境的复杂性。仿真实验表明,基于多奖惩标准的Q学习算法能够灵活地适应复杂变化的环境与状态,高效地完成学习任务。

关键词 Q学习算法,多奖惩标准,MAS,三维围捕
中图法分类号 TP181 **文献标识码** A

Research on Q-learning Algorithm Based on Multi-standard of Reward in MAS

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

Abstract Traditional Q-learning algorithm is based on a single standard of reward, when the environments and the state is changed, the single standard of reward may not be able to adapt to new environments and state in multi-agent system(MAS), instead, it may restrict the learning efficiency. This paper proposed a method of multi-agent Q-learning algorithm with multi-standard of reward. It adapt well to the changing environment and the state, complete the task in stages, different stages use different standards, so it can quickly complete the stage goal. In this paper, the simulation platform is pursuit problem in three-dimensional world. We increased the difficulty of rounding up and the complexity of the environment and state. Simulation results show that Q-learning algorithm based on multi-standard of reward can flexibly adapt to different environments and state, and efficiently complete learning tasks.

Keywords Q-learning algorithm, Multi-standard of reward, MAS, Pursuit problem in three-dimensional world

1 引言

增强学习算法是目前多智能体学习的研究热点之一。增强学习是基于动物学习心理学的有关原理,采用了人类和动物学习中的"尝试与失败"机制,强调在与环境的交互中学习,学习过程中要求获得评价性的反馈信号(reward/reinforcement signal,也称回报或增强信号),以极大化未来的回报为学习目标。增强学习由于不需要给定各种状态下的教师信号,因此在求解先验知识较少的复杂优化决策问题中具有广泛的应用前景[1]。Q学习算法是增强学习算法中的一类重要的学习算法,也是目前的研究热点。

传统的Q学习算法在整个学习过程中采用单一的奖惩标准,不考虑环境与状态的变化。许多研究者对奖惩标准做了一定的研究。范波、潘泉等提出引入中期目标,通过中期目标的奖惩信息来加速学习的方法[2]。陈宗海、段家庆等提出将奖惩分为动作奖惩和趋势奖惩,以此来提高学习效率[3]。这在简单的研究和应用中是可以的,但随着问题或环境的复杂化,单一奖惩标准的弊端也就体现了出来,无法做到"因地制宜、因时而异"。本文提出的多奖惩标准正是针对这一点,提出在不同状态下采用不同奖惩标准的方法,结合阶段目标与整体目标更为人性化地完成学习任务。

2 基于多奖惩标准的Q学习算法

2.1 单一奖惩标准的Q学习算法

Q学习算法是一种与模型无关的学习方法,它提供智能系统在马尔可夫环境中利用经历的动作序列选择最优动作的一种学习能力。Q学习基于的一个关键假设是智能体和环境的交互可看作一个Markov决策过程(MDP),即智能体当前所处状态和所选动作,决定一个固定的状转移概率分布、下一个状态,并得到一个即时回报。Q学习的目标是寻找一个策略可以最大化将来获得的报酬[4]。Markov决策过程由4个元组构成:环境有限状态集合S、有限动作空间集合A、状态转移函数$T:S\times A\to S$、回报函数$R:S\times A\to R$。

Q学习算法中的即时回报值r由预先定义的奖惩标准给出,在标准Q学习算法中奖惩标准单一且固定。在Q学习算法中,每个状态都对应一个$Q(s_t,a_t)$值,这个$Q(s_t,a_t)$值是按照所选择的策略持续执行而得到的累积回报,智能体在学习过程中根据这些$Q(s_t,a_t)$值选择动作。$Q(s_t,a_t)$值定义如下:

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;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。

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

式中，$s_t,s_t\in S$ 是当前所处状态，a_t 是按照某种策略所选择的动作，α 为学习率($\alpha>0$)，γ 为折扣因子($0\leqslant\gamma<1$)。其中的最优策略是 $\pi^*=\max Q(s_{t+1},a_{t+1})$，也就是选择具有最大回报值的动作。

2.2 多奖惩标准的Q学习算法

多智能体系统比单体复杂的地方在于多智能体系统中存在联合动作 $\vec{a}=(a^1,a^2,\cdots,a^n)$、联合状态 $\vec{s}=(s^1,s^2,\cdots,s^n)$ 以及联合奖惩 $\vec{r}=(r^1,r^2,\cdots,r^n)$。在简单的问题或智能体个数较少时并不会体现出其复杂性，但是随着问题的复杂化以及智能体数量的增加，从状态到动作的映射集合将会呈指数增长，也就是维数灾难问题。针对这一问题，本文采取的解决方法是让智能体制定目标，根据制定的目标时而独立地完成任务，时而合作完成任务。如果环境发生变化则重新制定目标，同时每个智能体有自己的 $Q(s_t,a_t)$ 表和共同的 $Q(s_t,a_t)$ 值表。当独立完成某项任务时则更新自己的 $Q(s_t,a_t)$ 表，需要合作完成任务时则更新共同的 $Q(s_t,a_t)$ 表。目标不同自然要求奖惩标准也不同。

多奖惩标准的Q学习算法与单一奖惩标准的区别就在于奖惩标准的增加。这些奖惩标准不是并行的，而是在某些特定的状态或是不同的环境中，算法会根据具体的情况选择适合这一阶段的奖惩标准，也就是所谓的“因地制宜、因时而异”。此时的回报值 r 可以表示为

$$r=\begin{cases} r_1=R_1, & \text{情境 } 1 \\ r_2=R_2, & \text{情境 } 2 \\ \vdots & \vdots \\ r_n=R_n, & \text{情境 } n \end{cases}$$

式中，R_i 表示第 i 种奖惩标准。此外上式中 r_i 的定义可仿照 r 的定义，即上式中的情境还可以细分，具体算法步骤如下：

步骤1　初始化 $Q(s_t,a_t)$ 表；

步骤2　观察 t 时刻的状态 s；

步骤3　按策略选择动作；

步骤4　执行所选动作，并观察下一个状态，判断所处的任务阶段和所处环境，选择适合该状态的奖惩标准，得到回报值 r；

步骤5　更新当前状态下所选用的 $Q(s_t,a_t)$ 表；

步骤6　判断是否满足终止条件，满足则停止学习，不满足则转步骤2。

3 基于多奖惩标准的围捕问题

3.1 三维空间中的围捕问题

多奖惩标准的Q学习算法是针对复杂环境的，本文将人工智能问题中经典的围捕问题作为仿真任务，并将其扩展到三维世界中，如图1所示。这是一个 7×7×7 的无边界格栅世界，猎物被分配在中心位置，如图1中的五角星所示。6个猎人被随机分配在格栅世界中，如图1中的实心点所示。图中每一个格栅代表一个状态，猎人可选择的动作有上、下、前、后、左、右6种。猎人不可同时到达同一位置，否则会被强行退回。6位猎人的任务各有不同，5号、6号猎人的目标为猎物上方和下方的位置，5号负责上面，6号负责下面。1号至4号猎人负责猎物前、后、左、右的位置。

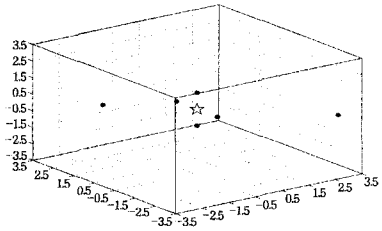

图1　围捕问题位置分布图

猎人分两阶段完成目标，在第一阶段中1号至4号猎人的目标为到达猎物所在层，5号猎人的目标为猎物上层，6号猎人的目标为猎物的下层。在这个阶段，1号至4号猎人共享一张 $Q(s_t,a_t)$ 表，以 $|z|$ 的变化作为奖惩标准，z 为三维世界中的 z 轴。当 $|z|$ 变大则给予负回报值，$|z|$ 不变则给予零回报值，$|z|$ 变小则给予正回报值。5号猎人独享一张 $Q(s_t,a_t)$ 表，以 $|z-1|$ 作为奖惩标准，6号猎人独享一张 $Q(s_t,a_t)$ 表，以 $|z|-1|$ 作为奖惩标准，奖惩规则均为值变大给予负回报值，值不变给与零回报值，值变小给予正回报值。

当完成了第一阶段的任务后，猎人们进入第二阶段。在这个阶段中，5号猎人的目标为猎物正上方的位置，6号猎人的目标为猎物正下方的位置，均以 $\sqrt{x^2+y^2}$ 的变化作为奖惩标准，x 代表三维世界中的横坐标，y 代表三维世界中的纵坐标。当 $\sqrt{x^2+y^2}$ 的值变大则给予负回报值，当 $\sqrt{x^2+y^2}$ 的值不变则给予零回报值，当 $\sqrt{x^2+y^2}$ 变小则给予正回报值。1号至4号猎人在第二阶段的目标需要自己选择。首先，猎物所在层会被分成4个区域，如图2所示，猎人会根据自己所处的位置与各坐标轴的夹角来判断自己的目标区域，进而确定目标区域中的目标位置。1号至4号猎人在第二阶段又会分成两小阶段，第一小阶段，猎人以形成包围圈为目标，即到达各自的目标区域，此时1号至4号猎人有各自的 $Q(s_t,a_t)$ 表，以和目标区域的坐标轴之间夹角的变化作为奖惩标准，夹角变大则给予负回报值，夹角不变则给予零回报值，夹角变小则给予正回报值。到达目标区域后就进入了第二小阶段，以缩小与猎物之间的距离即将 $\sqrt{x^2+y^2}$ 作为奖惩标准，奖惩规则也是 $\sqrt{x^2+y^2}$ 变大给予负回报值，$\sqrt{x^2+y^2}$ 不变则给予零回报值，$\sqrt{x^2+y^2}$ 变小则给予正回报值。此时1号至4号猎人更新的是以第二小阶段标准为依据的 $Q(s_t,a_t)$ 表。

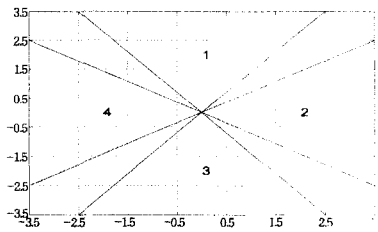

图2　目标区域图

整个围捕过程中，猎物在一个稍小的 5×5×5 的区域中随机游走，游走速度慢于猎人围捕的速度，猎物位置发生变化后，1号至4号猎人会再次协商，重新分配该阶段的目标。

3.2 仿真实验及结果分析

仿真实验参数设定如下：$\alpha=0.1$，$\gamma=0.95$，所有 $Q(s_t,a_t)$ 表的初值均为 0。猎人可移动的最大步数为 300，任一猎人移动超过 300 步则此次围捕失败。实验分两组，第一组为单一奖惩标准的 Q 学习算法，第二组为多奖惩标准的 Q 学习算法。每组实验分为 100 轮，每一轮实验均是在上一轮学习的基础上继续学习，每轮实验开始时猎人的位置都会被随机分配。

图 3 所示的是基于单奖惩标准的 Q 学习算法的实验结果，图 4 是基于多奖惩标准的 Q 学习算法的实验结果。将 100 轮实验分成 20 组，每组 5 轮，统计这 5 轮围捕中成功的次数，如图 3 和图 4 的左图所示，横坐标是组次，纵坐标是每组成功次数。同时将 100 轮实验中所有围捕成功的实验统计出来，求出每次成功围捕每个猎人行走的平均步数，如图 3 和图 4 的右图所示，横坐标是成功围捕的轮次，纵坐标是每个猎人行走的平均步数。

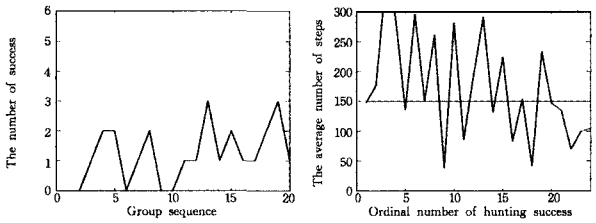

图 3 基于单奖惩标准的 Q 学习算法实验结果

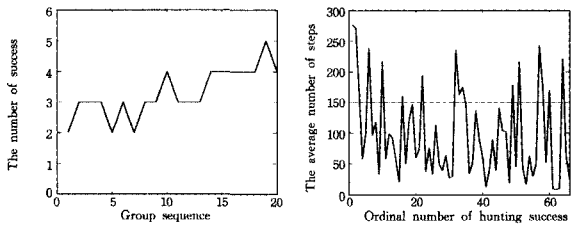

图 4 基于多奖惩标准的 Q 学习算法实验结果

从图 3 和图 4 的左图可以明显地看到，随着实验轮数的增加，基于多奖惩标准的 Q 学习算法的成功次数在明显增加，上升趋势也很明显。从图 3 和图 4 的右图可以更明显地看到，在 100 轮试验中基于单奖惩标准的 Q 学习算法只成功了 20 多次，而基于多奖惩标准的 Q 学习算法则成功了 60 多次。图 3 和图 4 的右图中有一条以 150 为基准的水平线，可以看到，基于单奖惩标准的 Q 学习算法在实验中每次成功围捕所需要行走的平均步数大部分在 150 步以上，而基于多奖惩标准的则大部分在 150 步以下。

由此可以看出，基于多奖惩标准的 Q 学习算法较基于单奖惩标准的 Q 学习算法，无论是在成功的次数上还是行走的平均步数上都有很明显的优势。

结束语 本文提出了基于多奖惩标准的 Q 学习算法，该算法是针对环境复杂、状态较多的学习场景而提出的。传统的单奖惩标准 Q 学习算法过于单一，无法灵活地适应环境或状态的变化，而多奖惩标准的 Q 学习算法减少了单一标准的束缚，避免了许多重复的工作，可以较灵活地适应不同的环境和状态。同时在学习过程中制定阶段目标，分段完成任务，真正做到“因地制宜、因时而异”。从结果上我们也可以明显看到，基于多奖惩标准的 Q 学习算法的成功次数是单奖惩标准 Q 学习算法的 2～3 倍，所需步数减少了近一半，整体性能也有很大的提高。因此，基于多奖惩标准的 Q 学习算法能够灵活适应动态环境，高效地完成学习任务。

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] 范波，潘泉，等. 多智能体学习中基于知识的强化函数设计方法[J]. 计算机工程与应用，2005，3：77-79

[3] 陈宗海，段家庆，等. 针对机器人觅食任务的强化学习算法及其仿真研究[C]// 系统仿真技术及其应用. 2008：252-256

[4] 宋清昆，胡子婴. 基于经验知识的 Q-学习算法[J]. 自动化技术与应用，2006，25(11)：10-12

[5] Notsu A, Ichihashi H. State and action space segmentation algorithm in Q-learning[C]// IEEE International joint conference on neural networks. 2008：2384-2389

[6] 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

[7] 李铁. 基于多 Agent 交互的团队学习仿真研究[D]. 山西：山西大学，2009

[8] 叶超群. 多 Agent 复杂系统分布仿真平台中的关键技术研究[D]. 长沙：国防科学技术大学，2006

[9] 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

[10] 胡子婴. 基于智能体系统的 Q-学习算法的研究与改进[D]. 哈尔滨：哈尔滨理工大学，2007

[11] Stone P, Veloso M. Multiagent Systems: A Survey from a Machine Learning Perspective[J]. Autonomous Robots 8, 2000：345-383

(上接第 219 页)

[3] Robert G, Gu Yun-hong. Data mining using high performance data clouds: experimental studies using sector and sphere [C]// Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008

[4] Grossmam R L, Gu Yum-homg, Michael S, et al. Compute and storage clouds using wide area high performance network [J]. Future Generation Computer Systems, 2009, 25：179-183

[5] Noordhuis P, Heijkoop M, Lazovik A. Mining twitter in the cloud: a case study [C]// Proceedings of the 2010 IEEE 3rd International Conference on Cloud Computing. 2010：107-114

[6] Piatetsky-Shapiro G. Knowledge discovery in databases: 10 years after[J]. SIGKDD Explorations, 2000, 1(2)：59-61

[7] 王鹏. 走进云计算[M]. 北京：人民邮电出版社，2009