ISSN 1000-1239/CN 11-1777/TP
50(2): 240-247, 2013

一种多动机强化学习框架

赵凤飞　　覃　征
(清华大学计算机科学与技术系　北京　100084)
(zhaofengfei@gmail.com)

A Multi-Motive Reinforcement Learning Framework

Zhao Fengfei and Qin Zheng
(*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

Abstract The traditional reinforcement learning methods such as Q-learning, maintain a table that maps the states to the actions. This simple dual-layer mapping structure has been widely used in many applied situations. However, dual-layer mapping structure of state-action lacks flexibility, while priori knowledge can not be effectively used to guide the learning process. To solve this problem, a new reinforcement learning framework is proposed, called multi-motive reinforcement learning (MMRL). Between state layer and action layer, MMRL framework introduces motive layer, in which multiple motives can be set based on experience. In this way, the original state-action dual-layer structure is extended to state-motive-action triple-layer structure. Under this framework, two new corresponding algorithms are presented, the first is MMQ-unique algorithm and the second is MMQ-voting algorithm. Moreover, it is stated that traditional reinforcement learning methods can be seen as a degenerate form of multi-motive reinforcement learning. That is to say, multi-motive reinforcement learning framework is a superset of traditional methods. This new framework and the corresponding algorithms improve the flexibility of reinforcement learning by adding the motive layer, and make use of priori knowledge to speed up the learning process. Experiments demonstrate that, multi-motive reinforcement learning can get better performance than the traditional reinforcement learning methods significantly by setting reasonable motives.

Key words reinforcement learning; multi-motive; Q learning; MMQ-unique algorithm; MMQ-voting algorithm

摘　要 以Q学习为代表的传统强化学习方法都是维持一个状态与动作的映射表.这种状态-动作的二层映射结构缺乏灵活性,同时不能有效地使用先验知识引导学习过程.为了解决这一问题,提出了一种基于多动机强化学习(MMRL)的框架.MMRL框架在状态与动作间引入动机层,将原有的状态-动作二层结构扩展为状态-动机-动作三层结构,可根据经验设置多个动机.通过动机的设定实现了先验知识的利用,进而加快了强化学习的进程,提高了强化学习的灵活性.实验表明,通过合理的动机设定,多动机强化学习的学习速度较传统强化学习有明显提升.

关键词 强化学习;多动机;Q学习;MMQ-unique算法;MMQ-voting算法

中图法分类号 TP181

收稿日期:2011-10-07;**修回日期**:2012-07-20

基金项目:国防"十一五"预研基金项目(402040202);国防"十二五"预研基金项目(041802008)

传统的强化学习算法,诸如 Q 学习[1]、Sarsa 学习[2-4]等都有一个共同的特点,就是要维护和更新一组状态到动作的映射,对这种映射的学习也是这些算法的主要任务.但这种从状态到动作的直接映射造成了状态空间和动作空间的高度耦合,没有为先验知识的引入留出空间,因此从结构上缺乏灵活性.传统的强化学习不需要先验知识.无需先验知识是强化学习的重要优点,但在另一方面也是强化学习的弱点.在传统强化学习应用的场合,通常是没有一个完整的先验知识域,但没有足够的知识并不等于"一无所知".而完全忽略领域知识只能造成强化学习算法的低效.本文提出了一种多动机强化学习方法(multi-motive reinforcement learning,MMRL),通过引入动机层,改变了传统的二层映射的方法,形成了从状态到动机、再从动机到动作的三层结构,使用者可以在动机层指定若干个动机,这些动机的设定表示了使用者希望学习 agent 进化的方向.多动机强化学习通过动机层的引入,有效地利用了先验知识,从而缩短了强化学习的学习进程.

Dietterich 于 1998 年提出了 MAXQ 分层强化学习方法[5-8],该方法基于半马尔可夫模型(semi-Markov decision process,SMDP),通过分层的方式实现了先验知识的引入,把整体的学习任务分为若干个可复用的子马尔可夫过程,从而加速了学习.但在实际问题中,很多决策过程难以人为地分解为若干子过程,例如在战斗机格斗行为的强化学习问题中,战斗机的各种机动动作不断重复出现,各种动作组合不断变换,很难准确地把格斗过程划分为若干子过程.多任务强化学习(multi-task reinforcement learning)是近年来机器学习领域的一个热点[9-13],该方法将多任务学习的思想引入强化学习领域,实现了用更少的样本数得到近似最优的结果.但本文所述的多动机与多任务是不同的,多任务强化学习是指通过识别相似结构实现使用少量样本对多个学习任务进行强化学习,而在多动机的强化学习中,动机是动作产生的背后原因,一个或多个动机的作用导致了某一动作的执行.通过多动机的引入改变了强化学习的学习进程,并且本文还说明了在没有先验知识来设定额外动机的情况下,多动机强化学习将退化为传统的强化学习,可见多动机强化学习是一种更广义、更灵活的强化学习框架.

本文针对现有强化学习方法的不足,提出了一种多动机强化学习框架,并给出了两种该框架下的算法:MMQ-unique 和 MMQ-voting.多动机强化学习框架通过引入动机层实现了先验知识的利用.本文通过经典的出租车问题对算法进行实验,对比了多动机强化学习与传统 Q 学习的性能,证明了多动机强化学习的效率.还通过战斗机 1v1 格斗仿真的实际应用展示了利用多动机强化学习框架设置不同动机时对仿真实验效果的影响.

1 强化学习

1.1 马尔可夫决策过程

马尔可夫决策过程(Markov decision process,MDP)是强化学习的理论基础[14-15].马尔可夫决策过程可用一个六元组$\langle S;A;P;R;\gamma;D\rangle$表示,其中 S 是过程的状态空间,A 是过程的动作空间,P 表示马尔可夫动作迁移模型,$P(s;a;s')$表示从状态 s 采取动作 a 到 s'的转移概率,R 代表回报函数,$R(s,a)$表示在状态 s 下执行动作 a 的预期回报,γ 是 0 到 1 之间的实数,是未来回报的折扣因子,D 代表初始的状态分布.对于一个 MDP,确定性的策略 π 是一个映射,$\pi:S\rightarrow A$,是从状态到动作的映射,$\pi(s)$表示在状态 s 下的动作选择.

1.2 强化学习

强化学习(reinforcement learning,RL)又称再励学习、增强学习,是机器学习的一个重要分支,其特点是通过与环境的交互学习来不断改进性能.传统的强化学习是学习从环境状态到行为的一组映射,以使得系统行为能从环境中获得最大的累积回报.和监督学习技术不同,强化学习不需要带标记的训练样例,而是通过不断试错的方式在交互中发现最优的动作策略[16-17].

在强化学习中,每当在某状态 s_t 下执行一动作 a_t,学习 agent 会收到一个实数回报 r_t,之后系统进入下一个状态.上述过程的不断重复产生了一系列的状态、动作和立即回报.学习 agent 的任务是学习一组策略 π:它使这些回报总和的期望值 $r_0+\gamma r_1+\gamma^2 r_2+\cdots$,$0\leqslant\gamma<1$ 最大化,其中后面的回报值随折扣因子 γ 减少.定义 $V^{\pi}(S_t)=r_t+\gamma r_{t+1}+\gamma^2 r_{t+2}+\cdots=\sum_{i=0}^{\infty}\gamma^i r_{t+i}$,学习的目标就是使 $V^{\pi}(S_t)$ 最大化.那么最优策略可以表示为

$$\pi^{*}=\arg\max_{\pi}V^{\pi}(s),\ (\forall s). \tag{1}$$

1.3 Q 学习

Q 学习算法由 Watkins 等人首先提出[18].Q 学习的一大特点是模型无关性,不需要有状态转移函数

和动作函数的相关知识,而是完全从环境中学习.根据动态规划的思想,在当前状态 s 下最好的动作 a 能够最大化当前回报 $r(s, a)$ 与后继状态的最优值函数 V^* 的和,也就是:

$$\pi^*(s)=\arg\max_a[r(s,a)+\gamma V^*(\delta(s,a))], \quad (2)$$

其中 $\delta(s, a)$ 表示执行动作 a 后的后继状态.在实际问题中,值函数 V 通常是难以直接学习的,所以引入了 Q 函数:

$$Q(s,a)=r(s,a)+\gamma V^*(\delta(s,a)). \quad (3)$$

状态 s 下的最优策略可以表示为

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

学习 agent 执行动作得到立即回报,然后根据下面公式对 Q 值进行更新:

$$Q(s,a)\leftarrow Q(s,a)+\alpha[r+\gamma\max_{a'}Q(s',a')-Q(s,a)], \quad (5)$$

其中 α 表示每次学习的更新率.

由于 Q 学习环境无关的特性,它被公认为是最有效的强化学习算法之一,同时也是目前应用最为广泛的强化学习算法.

2 多动机强化学习

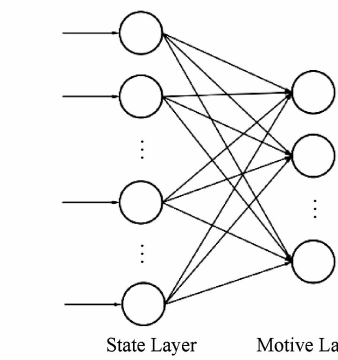
2.1 从二层结构到三层结构

在这里,我们首先对马尔可夫决策过程进行扩展,得到动机马尔可夫决策过程(motive Markov decision process, MMDP),该过程可以被表示为七元组 $\langle S;M;A;P;R;\gamma;D\rangle$,其中 S 是过程的状态空间, M 是过程的动机空间, A 是过程的动作空间, P 表示马尔可夫动作迁移模型, $P(s;a;s')$ 表示从状态 s 采取动作 a 到 s' 的转移概率, R 代表回报函数, $R(s, a)$ 表示在状态 s 下执行动作 a 的预期回报, γ 是 0 到 1 之间的实数,是未来回报的折扣因子, D 代表初始的状态分布.对于一个 MMDP,确定性的策略 π 可以表示为 $\pi=\pi_1\pi_2$,其中 $\pi_1: S\rightarrow M$ 是从状态到动机的映射; $\pi_2: M\rightarrow A$ 是从动机到动作的映射. π_2 可以根据实际情况,通过 IF-THEN 规则指定. $\pi_2(\pi_1(s))$ 表示在状态 s 下的动作选择.也就是说,在 MMDP 中,动作选择是分成两步完成的,第 1 步是由状态选择动机,第 2 步是由动机生成动作.

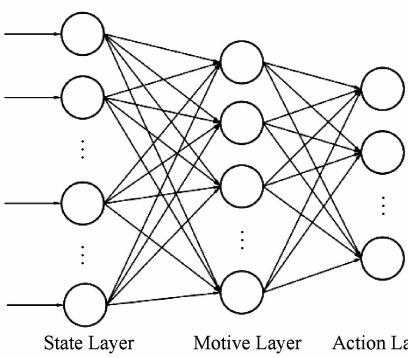
在一般的强化学习问题中,动机的设定通常可以采用以下形式:原问题动作集中的所有动作都对应一种基本动机;在基本动机之外,根据相应问题的先验知识和偏好再设置若干个复杂动机,这些动机与动作之间不是简单的直接对应关系,而是一种更复杂的对应,这种对应通过 IF-THEN 语句来加以说明.例如战斗机格斗仿真中,使飞机的每一种基本动作如拉起、俯冲、加速、减速等都对应一种动机,同时根据需要可以再设置若干动机,如设置主动接敌战斗动机和自我保全动机,这些动机在学习过程中会选择适合的动作以实现动机的设定意图.

多动机强化学习通过引入动机层,将原有的二层映射转化为三层映射.如图 1 所示,与状态直接映射的不再是动作,而是动机.多动机强化学习框架扩展了 MDP,改变了动作选择的流程,在动作选择的流程中增加了一个环节从而实现了先验知识的引入.

在传统的 Q 学习中,状态 s 与动作 a 直接建立联系,学习过程是对 $Q(s,a)$ 值的学习.而多动机强化学习是对 $Q(s,m)$ 值的学习, $Q(s,m)$ 值的大小表示在状态 s 下动机 m 的强度.本文提出了两种算法来利用 $Q(s,m)$ 进行动作选择并对 Q 值进行更新,它们分别是 MMQ-unique 算法和 MMQ-voting 算法.



(a) Traditional dual-layer structure



(b) Triple-layer structure of MMRL

Fig. 1 From dual-layer structure to triple-layer structure.

图 1 从二层结构到三层结构

2.2 MMQ-unique 算法

MMQ-unique 算法在每次动作选择中发挥作用的动机是唯一的,这个动机按照一定的原则依概率

产生,动机的选择方式与传统强化学习动作的选择方式类似,可以直接复用.这个唯一的动机 m 产生相应的动作 a 并得到环境给予的回报 r.在每一个时间步中,r 只用来更新动机 m 所对应的 $Q(s,m)$值.

在 MMQ-unique 算法中,式(3)中的 Q 函数的形式变为

$$Q(s,m) = r(s,\pi_2(m)) + \gamma V^*(\delta(s,\pi_2(m))), \tag{6}$$

其中 $a=\pi_2(m)$,此时状态 s 下的最优策略可以表示为

$$\pi^*(s) = \arg\max_a Q(s,\pi_2^{-1}(a)). \tag{7}$$

式(5)中的 Q 值更新公式改写成:

$$Q(s,m) \leftarrow Q(s,m) + \alpha[r + \gamma \max_{a'} Q(s',m') - Q(s,m)]. \tag{8}$$

MMQ-unique 算法的伪代码如算法 1 所示.

算法 1. MMQ-unique 算法.

① Set initial $Q(s,m)$ value arbitrarily;
② Repeat (for each episode)
③ Set initial s;
④ Repeat (for each step of episode)
⑤ Choose m from s using policy $m=\pi_1(s)$;
⑥ Choose a from m using policy $a=\pi_2(m)=\pi_2(\pi_1(s))$;
⑦ Take action a, and observe r, s';
⑧ $Q(s,m)\leftarrow Q(s,m)+\alpha[r+\gamma\max_{a'}Q(s',m')-Q(s,m)]$;
⑨ $s\leftarrow s'$;
⑩ Until the episode ends;
⑪ Until all episodes finish.

2.3 MMQ-voting 算法

和 MMQ-unique 算法类似,MMQ-voting 算法也是对 $Q(s,m)$进行学习,二者的区别在于:在每一次动作选择过程中 MMQ-voting 算法并不是由状态 s 选择唯一的动机 m,而是通过多个动机对应的 Q 值按比例加权综合来选择动作 a.正如人类的决策过程一样,一个合理决策经常是各种动机综合考虑的结果,在 MMQ-voting 算法中,我们也采用了这种方式,称之为"投票机制".a 执行后得到的回报 r 也同时给予所有对 a 投票的动机,进而更新这些相应动机的 $Q(s,m)$值.

我们定义 $w(Q)$为每个动机选择动作的投票权重,此时每个动作 a 的得票可以表示为

$$vote(a) = \sum_{a=\pi_2(m_i)} w(Q(s,m_i)), \tag{9}$$

其中得票最多的动作胜出并被执行.投票机制的本质是在动作选择过程中实现学习 agent 多种动机的折中,而 $Q(s,m)$值的大小决定了每个动机在动作选择问题上的"发言权".在 MMQ-voting 算法中,式(3)中的 Q 函数的形式变为

$$Q(s,m) = r(s,\arg\max_a vote(a)) + \gamma V^*(\delta(s,\arg\max_a vote(a))), \tag{10}$$

其中$\arg\max_a vote(a)$表示投票选出的动作,此时状态 s 下的一种近似最优策略可以表示为

$$\pi^*(s) = \arg\max_a vote(a). \tag{11}$$

MMQ-voting 算法的 Q 值更新公式同样是式(8),但不同的是对每一个在投票中获胜的动机 m 都给予回报 r,所以一个时间步中将可能有多个 Q 值被更新.MMQ-voting 算法的伪代码如算法 2 所示.

算法 2. MMQ-voting 算法.

① Set initial $Q(s,m)$ value arbitrarily;
② Repeat (for each episode)
③ Set initial s;
④ Repeat (for each step of episode)
⑤ Calculate $vote(a)$ for each a using $vote(a) = \sum_{a=\pi_2(m)} w(Q(s,m))$;
⑥ Choose a with the max $vote(a)$;
⑦ Take action a, and observe r, s';
⑧ Repeat (for each m that votes a)
⑨ $Q(s,m)\leftarrow Q(s,m)+\alpha[r+\gamma\max_{a'}Q(s',m')-Q(s,m)]$;
⑩ $s\leftarrow s'$;
⑪ Until all motives of a finish;
⑫ Until the episode ends;
⑬ Until all episodes finish.

在上述两种算法中都有可能会出现动机与动机之间相互矛盾的情况,导致强化学习具有不同的动作选择倾向.在 MMQ-unique 算法中,这种相矛盾动机的选择具有排他性,仅有一个动机会被最终选择来发挥作用.而在 MMQ-voting 算法中相互矛盾的动机会同时在投票中发挥作用,每个动机发挥作用的大小取决于投票权重函数 $w(Q)$的设定.投票权重函数的形式也体现了先验知识的作用.

2.4 MMRL 框架与传统强化学习的关系

普通的强化学习方法可以作为多动机强化学习的一种退化形式.因为当每个动作与每个动机直接对应时,动机层到动作层的映射通过 IF-THEN 表示设定为 IF:当前动机是动作 a 对应的动机,THEN:

选择动作 a 执行. 在这种设定下,多动机强化学习就退化为传统的二层映射的强化学习. 这说明,多动机强化学习是比普通学习算法更广义的强化学习框架,当且仅当先验知识为零,无法设置任何其他动机时,才会退化为普通强化学习方法.

多动机强化学习框架不仅适用于表格式的强化学习,函数近似的强化学习同样可以使用,只需在状态层与动作层间的映射中引入函数近似. 此时,$S\rightarrow A$ 的函数估计变为 $S\rightarrow M$ 的函数估计.

3 实验结果

3.1 出租车问题

让我们通过一个经典的、简单的例子来验证我们的方法. 图 2 中显示了一个 5×5 的含有出租车 agent 的方格世界,在此环境中有 4 个特殊的地点,分别标注为 R(red), B(blue), G(green)和 Y(yellow)[19-20]. 出租车问题采用插曲式(episodic)的强化学习. 在每一个插曲中,出租车随机地选择一个地点作为起点,一位乘客随机地出现在 4 个特殊位置之一. 这个乘客要到达这 4 个地点中的某一个(也是随机选出的). 出租车必须到乘客所在的位置接载,然后将乘客送到目的地卸载. 为了统一描述,如果乘客本身已在目的地,也要有上车和下车的过程.

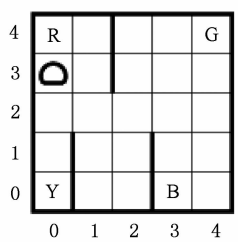


Fig. 2 The taxi domain.

图 2 出租车问题示意图

每一个插曲以乘客到达目的地结束. 在整个过程中,一共有 6 个可选动作,分别是导航动作东、西、南、北,以及上车和下车动作. 每一个普通动作有−1 的奖赏,当到达目的地时得到+20 的奖赏,在错误的时机上车、下车有−10 的回报. 导致撞墙的导航动作没有实际状态改变,但给予回报−1. 所有动作的执行结果是确定性的.

我们通过实验将 MMQ-unique 算法、MMQ-voting 算法与传统的 Q 学习算法进行比较. 实验数据为重复 100 次实验的平均值. 在第 1 组实验中,按照 2.1 节中所述的动机设定的方式,共设置 9 个动机,前 6 个分别为东、西、南、北,以及上车和下车,分别与 6 个基本动作相对应;第 7 到第 9 个动机通过先验知识设定,以引导学习器加快学习过程,分别为:靠近乘客、靠近终点和避免引起撞墙的导航. 这些动机与动作之间关系的 IF-THEN 表达如表 1 所示:

Table 1 IF-THEN Mapping In Taxi Domain

表 1 出租车问题的 IF-THEN 映射

No.	IF Statement	THEN Statement
1	m==East	a= East
2	m==West	a=West
3	m==South	a=South
4	m==North	a=North
5	m==Pickup	a=Pickup
6	m==Putdown	a=Putdown
7	m==Move to passenger	a=Choose an action that shorten the distance with passenger randomly
8	m==Move to destination	a=Choose an action that shorten the distance with destination randomly
9	m==Don't hit wall	a= Choose an action that won't hit the wall randomly

我们限定学习 agent 学习 1 000 个周期,学习率 $\alpha=0.2$,折扣因子 $\gamma=0.95$,MMQ-voting 算法的投票权重函数 $w(Q)$设置为

$$w(Q(s,m)) = e^{\lambda\cdot Q(s,m)}. \tag{12}$$

实验中 λ 取值为 3,图 3 给出了实验结果,横坐标表示学习的周期数,纵坐标表示学习 agent 执行的总步数(动作总数). 从图 3 可以看出,MMQ-unique 和 MMQ-voting 两种算法的性能均超越了传统 Q 学习,在完成相同学习周期数的情况下,所用的动作数均有减少,其中 MMQ-voting 算法的性能尤为突出,可见投票机制是选择动作的有效方式. 而 MMQ-

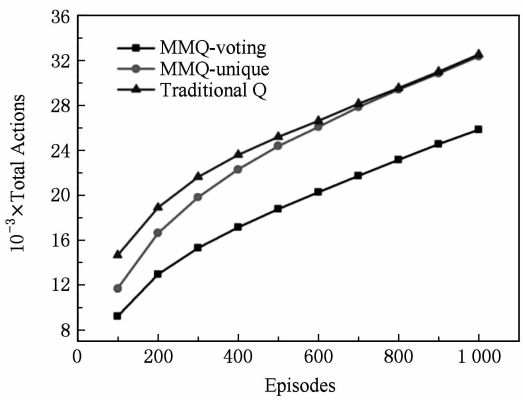


Fig. 3 Experiments result (9 motives, 1 000 episodes).

图 3 9 个动机学习 1 000 周期的结果

unique 算法的优势主要体现在初始阶段.

为说明当设置不同的动机时多动机强化学习算法的表现是有差别的,我们以 MMQ-voting 算法为例,进行第 2 组实验,观察其在两种动机设置下的学习效果. 在第 2 组实验中,共设置 8 个动机,与第 1 组实验的前 8 个动机完全相同. 各个参数和投票权重函数的选择也均与第 1 组实验相同.

表 2 给出了 9 个动机时和 8 个动机时 MMQ-voting 算法实验结果的对比,从表 2 可以看出,MMQ-voting 算法在 8 个动机的情况下学习效率更高,说明第 2 组实验中的动机设定更为合理. 由于动机的选择对学习的结果有着明显的影响,合理设置动机是多动机强化学习达到良好性能的重要因素.

Table 2 Comparisons for MMQ-voting Algorithm in Two Experiments

表 2 MMQ-voting 算法在两组实验中的结果比较

Episodes	Mean Total Actions	
	9 motives	8 motives
100	9 223.67	9 152.01
200	12 958.84	12 796.22
300	15 314.84	15 167.45
400	17 132.07	16 992.16
500	18 760.85	18 583.85
600	20 259.51	20 126.61
700	21 735.27	21 480.02
800	23 151.22	22 940.61
900	24 557.65	24 256.17
1 000	25 836.61	25 626.02

3.2 战斗机 1v1 格斗仿真

本节我们将通过一个复杂的实际应用来说明多动机强化学习中不同动机对学习效果的影响.

战斗机空战仿真是一个较为复杂的建模过程. 通过强化学习让计算机仿真系统中的战斗机自我学习,通过试错来达到接近人类飞行员的作战方式是一个充满挑战的问题. 我们采用值函数近似的强化学习来对 8 个特征进行学习,这些特征分别是偏离角、脱离角、相对距离、速度矢量夹角、速度平方、速度平方差、飞行高度和高度差. 选取的特征的具体含义参见文献[21].

本实验为战斗机设定 6 个基本飞行动作,分别是水平加速、水平减速、左转弯、右转弯、拉起和俯冲. 在尾随敌机情况下,当与敌机距离小于 10 km、我方速度矢量与敌机夹角小于 30°并且双方速度矢量夹角小于 40°时可以发起攻击,实验中假定每次攻击的命中率为 70%,击中敌机可以获得正回报,被击中的一方获得负回报.

我们设定 8 个动机,按照 2.1 节中所述的动机设定的方式,前 6 个动机分别与 6 个基本动作对应,另外 2 个动机根据作战偏好,分别设定为主动接敌战斗和自我保全. 动机与动作之间关系如表 3 所示:

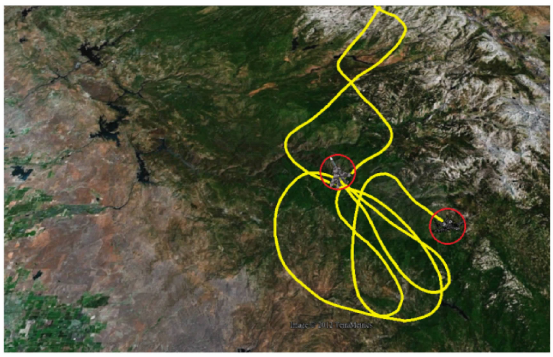
Table 3 IF-THEN Mapping in Air Combat

表 3 战斗机空战仿真中的 IF-THEN 映射

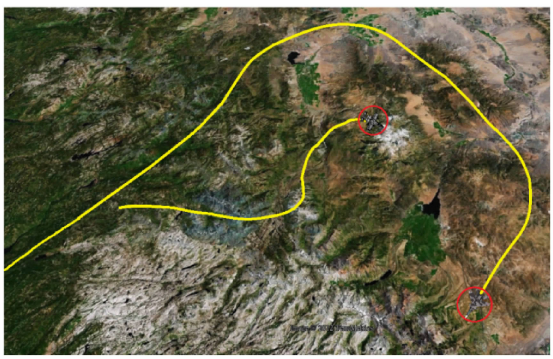
No.	IF Statement	THEN Statement
1	m==Horizontally accelerate	a=Horizontally accelerate
2	m==Horizontally decelerate	a=Horizontally decelerate
3	m==Turn left	a=Turn left
4	m==Turn right	a=Turn right
5	m==Pull up	a=Pull up
6	m==Dive	a=Dive
7	m==Initiative to combat	a=Point to the enemy and narrow the difference in height
8	m==Protect itself	a=Choose actions that keep itself away from the state being attacked

我们对 MMQ-voting 算法稍作改变,使得 $Q(s, m)$的查找和更新并不使用表格方式,而是使用值函数近似. 使用修改后的 MMQ-voting 算法进行 2 次模拟实验,在第 1 次模拟中,我们在投票权重函数 $w(Q)$中为主动接敌战斗动机设置了较高的权重,使其发挥主要作用,使战斗机更为好斗. 在第 2 次模拟中,我们在投票权重函数 $w(Q)$中为自我保全动机设置了较高的权重,使战斗机更为谨慎和倾向于自保.

如图 4 所示,曲线代表飞行轨迹,曲线末端的圆圈标出了当前两架飞机的位置. 两组实验中均得到类似于人类飞行员的飞行轨迹,而动机的设定明显地影响了战斗机的作战行为,在图 4(a)中,两架战斗机不断地主动向对方发起战斗,飞行轨迹相互纠缠. 而图 4(b)中战斗机且打且避,飞行轨迹并没有紧密地纠缠在一起. 仿真表明,不同动机的设定意图在学习过程中得到了体现,多动机强化学习的这种特点尤其适用于战斗仿真这一类需要根据战略意图改变行为方式的应用.



(a) Encourage the combat motive



(b) Encourage the preservation motive

Fig. 4 Air combat simulation result.

图 4 战斗机空战仿真结果

4 总 结

多动机强化学习在状态与动作之间引入了动机层,在结构上增加了强化学习的灵活性,通过动机层的引入,也使得人们的先验知识能够以一种偏好的形式指导学习过程.在实际应用中,人们完全没有任何先验知识的领域是少数的,很多情况下强化学习没有必要从零学起.实验表明,多动机强化学习框架及相应算法的引入,改善了强化学习的性能,加速了学习进程.并且多动机强化学习框架是对传统强化学习的推广,在没有先验知识来设定额外动机的情况下退化成传统的强化学习,这也保证了该框架对传统强化学习方法的兼容性.与 MAXQ 分层强化学习一样,MMRL 是一个强化学习的框架,而框架采用的具体算法可以替换或改进.在进一步的工作中,如下几个方向值得继续探索:

1) 如何设计一种动机的自动归纳方法以辅助使用者设置合理的动机;

2) 如何将多动机强化学习推广到多 agent 环境,使得多 agent 系统中带有不同动机的 agent 间可以交流信息,进行协作;

3) 多动机强化学习算法收敛性的严格理论证明还需要进一步的研究.

参 考 文 献

[1] Tsitsiklis J N. Asynchronous stochastic approximation and *Q*-learning [J]. Machine Learning, 1994, 16(3): 185-202

[2] Sutton R S. Generalization in reinforcement learning: Successful examples using sparse coarse coding [J]. Advances in Neural Information Processing Systems, 1996, 8: 1038-1044

[3] Singh S P, Stton R S. Reinforcement learning with replacing eligibility traces [J]. Machine Learning, 1996, 22(1/2/3): 123-158

[4] Sprague N, Ballard D. Multiple-goal reinforcement learning with modular Sarsa(0)[C] // Proc of the 18th Int Joint Conf on Artificial Intelligence. San Francisco: Morgan Kaufmann, 2003: 1445-1447

[5] Dietterich T G. The MAXQ method for hierarchical reinforcement learning [C] // Proc of the 15th Annual Int Conf on Machine Learning. San Francisco: Morgan Kaufmann, 1998: 118-126

[6] Barto A G, Mahadevan S. Recent advances in hierarchical reinforcement learning [J]. Discrete Event Dynamic Systems: Theory and Applications, 2003, 13(4): 41-77

[7] Shi Chuan, Shi Zhongzhi, Wang Maoguang. Online hierarchical reinforcement learning based on path-matching [J]. Journal of Computer Research and Development, 2008, 45(9): 1470-1476 (in Chinese)
(石川, 史忠植, 王茂光. 基于路径匹配的在线分层强化学习方法[J]. 计算机研究与发展, 2008, 45(9): 1470-1476)

[8] Zang Peng, Zhou Peng, Minnen D, et al. Discovering options from example trajectories [C] // Proc of the 26th Annual Int Conf on Machine Learning. New York: ACM, 2009: 1217-1224

[9] Lazaric A, Ghavamzadeh M. Bayesian multi-task reinforcement learning [C] // Proc of the 27th Annual Int Conf on Machine Learning. New York: ACM, 2010: 599-606

[10] Caruana R. Reinforcement learning with Gaussian processes [J]. Machine Learning, 1997, 28(1): 41-75

[11] Baxter J. A model of inductive bias learning [J]. Journal of Artificial Intelligence Research, 2000, 12: 149-198

[12] Xue Ya, Liao Xuejun, Carin L. Multi-task learning for classification with Dirichlet process priors [J]. Journal of Machine Learning Research, 2007, 8(1): 35-63

[13] Yu Kai, Tresp V, Schwaighofer A. Learning Gaussian processes from multiple tasks [C] // Proc of the 22nd Annual Int Conf on Machine Learning. New York: ACM, 2005: 1017-1024

[14] Zico K J, Ng A Y. Near-Bayesian exploration in polynomial time[C] // Proc of the 26th Annual Int Conf on Machine Learning. New York: ACM, 2009: 513-520

[15] Jason P, Lagoudakis M G. Binary action search for learning continuous-action control policies [C] // Proc of the 26th Annual Int Conf on Machine Learning. New York: ACM, 2009: 793-800

[16] Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey [J]. Journal of Artificial Intelligence Research, 1996, 4: 237-285

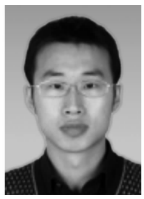
[17] Sutton R S, Barto A G. Reinforcement Learning [M]. Cambridge: MIT Press, 1998

[18] Watkins C J C H, Dayan P. Technical note. Q-learning [J]. Machine Learning, 1992, 8(3): 279-292

[19] Dietterich T G. Hierarchical reinforcement learning with the MAXQ value function decomposition [J]. Journal of Artificial Intelligence Research, 2000, 13: 227-303

[20] Jong N K, Stone P. Hierarchical model-based reinforcement learning: R-MAX+MAXQ [C] // Proc of the 25th Annual Int Conf on Machine Learning. New York: ACM, 2008: 432-439

[21] Ma Yaofei, Gong Guanghong, Peng Xiaoyuan. Cognition behavior model for air combat based on reinforcement learning [J]. Journal of Beijing University of Aeronautics and Astronautics, 2010, 36(4): 379-383 (in Chinese)
(马耀飞, 龚光红, 彭晓源. 基于强化学习的航空兵认知行为模型[J]. 北京航空航天大学学报, 2010, 36(4): 379-383



Zhao Fengfei, born in 1986. PhD. Student member of China Computer Federation. His main research interests include reinforcement learning and multi-agent system.

Qin Zheng, born in 1956. Professor and PhD supervisor. His main research interests include software architecture, information processing, E-commerce, and machine learning.