第16卷 第8期
2006年8月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.16 No.8
Aug. 2006

一类基于启发式搜索的激励学习算法

唐中勇,付 强,卓 佳,陈焕文
(长沙理工大学 计算机通讯工程学院,湖南 长沙 410076)

摘 要:激励学习已被证明是在控制领域中一种可行的新方法。相比其他的方法,它能较好地处理未知环境问题,但它仍然不是一种有效的方法。幸运的是,在现实世界中,智能体总是会有一些环境的先验知识,这些能形成启发式信息。启发式搜索是一种常用的搜索方法,有很快的搜索速度,但需要精确的启发式信息,这在有些时候难以得到。文中分析比较了启发式搜索和激励学习的各自特点,提出一类新的基于启发式搜索的激励学习算法,初步的实验结果显示了较好的性能。

关键词:启发式搜索;激励学习;启发式SARSA
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2006)08-0041-03

A Class of Reinforcement Learning Algorithm Based on Heuristic Search

TANG Zhong-yong, FU Qiang, ZHUO Jia, CHEN Huan-wen
(Dept. of Computer and Communication, Changsha Univ. of Sci. and Techn., Changsha 410076, China)

Abstract: The reinforcement learning has been proved to be a new applicable method in control field. It can solve the problems of unknown environment better than the others. But it isn't a very effective method yet. Fortunately in real world, the agent often has some knowledge of the environment, which can be used as heuristic information. The heuristic search is a very effective search method, which can search very quickly. But it need very precise heuristic information, which may be hard to get in complex environment. The characteristics of heuristic search and reinforcement learning are compared and a class of reinforcement learning algorithm on heuristic search is introduced. The preliminary empirical result shows better than the previous.

Key words: heuristic search; reinforcement learning; H-SARSA

0 引 言

在人工智能的搜索方法中,启发式搜索是一类重要的搜索方法。它通过评估函数来计算代价,以寻找最优的搜索方法[1]。比如,A*算法就是一种典型的启发式搜索算法,被普遍应用于博弈、电脑游戏的AI设计等等。启发式算法在合理的评估函数下有着相当优秀的性能。但是,启发式搜索需要设置精确的评估函数,这在环境未知的复杂状态下是很难做到的。激励学习通过奖赏函数,让智能体在与环境的交互中自行判断动作的优劣,它的优点在于无须事先知道环境模型,适合于在线学习,已证明在电梯调度[2]、作业调度[3]、游戏、机器人导航等方面是一种有效的、实用的方法。但是,由于激励学习理论上需要通过遍历整个状态空间(虽然实际应用中大部分激励学习算法并不这么做),才能保证收敛于最优解,故这并不是一种高效的算法。因此,假设在一个较为复杂的环境下,智能体只知道很少的环境信息,换句话说,只能设置简单的、不精确的评估函数。笔者尝试将利用启发式搜索中的评估函数与激励学习算法结合起来,通过实验对比可知,这样可以取得不错的效果。

收稿日期:2005-11-16
作者简介:唐中勇(1977-),男,湖南衡阳人,硕士研究生,研究方向为激励学习;陈焕文,博士,教授,研究方向为激励学习、人工智能等。

1 激励学习及其算法

激励学习是最近研究较多的一种机器学习方法。激励学习和传统的监督学习的不同在于:在激励学习中,智能体并不被告之哪个动作是最好的,而是让它自己与环境交互,不断地试错(trial and error),从环境中得到奖惩信息,积累经验,然后让它自己判断哪个动作是最好的。这也是激励学习最吸引人的地方,因为在一个复杂的决策系统中,可能事先关于环境的信息是很少的,这样监督学习是不可行的,因此采用激励学习这种不断试错,从而自己学习的方法是必要的。激励学习已被证明在很多方面取得了不错的成果。

文中将利用一种在线学习的激励学习算法SARSA。SARSA算法最初由Rummery和Niranjan在1994年提出[4],它由传统的Q学习发展而来,是一种典型的TD(Temporal Difference)算法。TD算法结合了动态规划算法和Monte Carlo算法,既可以象Monte Carlo算法那样不需要模型,也可以象DP算法那样不需要等待最终的结果

就可以对值函数进行估计计算,文中将采用它进行实验。Q学习是最常用的激励学习算法之一[5],它是一种离线(off-line)的瞬间差分算法。SARSA学习是在Q学习的基础上提出的,它的更新规则与Q学习极为相似,不同之处在于 $Q'(s_{a+1},a_{t+1})$ 中的估计下一步动作 a_{t+1} 的选择,Q学习总是采取贪婪策略,而SARSA采取的是实际的探索策略,这种实际的探索策略采用随机概率的形式,让智能体随机选取动作,Q 值大的状态赋予的概率相对较大,而 Q 值小的状态赋予的概率相对较小。采用这种随机动作选择的策略在一定程度上可以解决Q学习所带来的易陷入局部最优的问题。它的更新准则如下:

$$Q(s_{t+1},a_{t+1}) = Q(s_t,a_t) + a[r_{t+1} + \gamma * Q(s_{a+1},a_{t+1}) - Q(s_t,a_t)]$$

这样,实际采取的动作 a'_{t+1} 总是与估计动作 a_{t+1} 相同,故称之为在线(on-line)算法。为了加快SARSA学习的学习速度,有学者提出了加入可行性跟踪(eligibility traces)的技术[6],称为SARSA(λ)。SARSA(λ)算法描述如下:

初始化:所有的 $Q(s,a) = 0, e(s,a) = 0$。

每次尝试(trial)重复下面的过程:

初始化 s,a

对每一次尝试的每一步(step)重复下面的算法:

(1) 选择动作 a,观测奖赏值 r 和后继状态 s';

(2) 根据当前的 Q 值表,采用特定的策略,从 s' 的动作列表中选择 a'。

$$\delta = r + \gamma \max Q(s',a') - Q(s,a) e(s,a) = e(s,a) + 1$$

δ 为时间差分值,γ 为折扣因子,$e(s,a)$ 为跟踪值。

(3) 对于所有的 $s,a, Q(s,a) = Q(s,a) + a\delta e(s,a)$,如果 $a \neq a', s = s'$,则 $e(s,a) = 0$;如果 $a = a', s = s'$,则 $e(s,a) = \lambda\gamma e(s,a)$。

λ 为有效性跟踪截断系数。

(4) 如果实验的步数超过了允许的最大步数或者 s' 为终点状态,则结束这次循环,进行下一次训练尝试,否则 $s = s', a = a'$,继续上面的过程。

(5)如果尝试的总次数超过了允许的次数,或者稳定地达到期望的控制目标,则结束学习过程。

2 基于启发式的激励学习

在人工智能领域,搜索的方法一般可以分为两类:盲目式搜索和启发式搜索。盲目搜索不利用环境信息,按照一定的规则对所有的状态空间进行遍历搜索,常用的有广度优先搜索和宽度优先搜索;而启发式搜索可以在搜索的同时利用环境信息作为启发信息,如在网络中,可以从上一级路由器中找到相应的路由表来确定下一步搜索的路线,加速问题的求解过程。在搜索过程中,关键的一步是如何选择下一个要考察的结点,如果在选择结点时如能充分利用与问题有关的特征信息,估计结点的重要性,就能在搜索时候选择重要性较高的结点,以利于求最优解,这个估计结点重要性的函数就称为评估函数。它定义为从初始结点 S_0 出发,约束地经过结点 x 到达目标结点 S_r 的所有路径中最小路径代价的估计值。其一般形式为:$f(x) = g(x) + h(x)$。其中,$g(x)$ 表示从初始结点 S_0 到达结点 x 的实际代价;$h(x)$ 代表从 x 到目标结点 S_r 的最优路径的评估代价,它体现了问题的启发式信息,其形式要根据问题的特性确定,$h(x)$ 称为启发式函数。一般而言,启发式搜索可以获得比盲目搜索快得多的效果,在实际中,它大量应用于电脑博弈、网络路由选择等等方面。对于启发式搜索而言,启发信息越多,越能加快搜索速度,在实际算法中,人们一般采用启发式函数来获得启发信息,这样,启发式函数设计的好坏就决定了搜索的效率。但是,在现实生活中的有些问题,环境是非常复杂而且是事先不可预知的,如移动机器人导航。这样,精确的评估函数是很难事先确定好的。因此,采用启发式搜索并不能保证搜索的高效。上文提过,激励学习适合于处理事先不知道环境模型的情况。因此,文中尝试将二者结合起来,提出了一种基于启发式的激励学习方法:H-SARSA。

这里从两种方法的特点入手。对于激励学习而言,最主要的问题是:一般假设智能体在求解问题时除了事先设定的奖惩函数,对环境毫无所知,这虽然是它的优势,但并不符合现实情况,同时,这也是效率不高的主要原因。而对于启发式搜索而言,如果没有精确的启发式函数,是很难保证求得最优解,同时,对于一些简单的启发式信息,以爬山法为例,是很容易陷入局部最优解的情况。但是,相对于激励学习而言,求解初期就算是简单的启发式信息也能引导智能体加快搜索的速度。基于这样的考虑,可以在搜索的初期以启发式为主,在智能体获得一定的经验后再以激励学习为主,辅以启发式信息。具体思路如下:设置一个总控制器 C,用来决定智能体选取下一步动作,其中可以包含若干个子控制器(至少含有一个启发式搜索策略,文中简化为两个:一个为启发式,另一个为SARSA),每个控制器具有不同搜索策略,可以按各自的策略分别选取智能体的下一步动作。给每个子控制器添加两个附加的参数变量:优先权 M 和选择概率 P,在文中为 M_s 和 M_h,分别为SARSA控制器和启发式控制器的优先级,P_s 和 P_h 分别为SARSA控制器和启发式控制器的动作选择概率。总控制器 C 将根据子控制器的这两个变量的值进行选择,首先比较优先级,如果相同再比较选择概率,值较大的控制器所决定动作将被总控制器所选择,作为智能体的下一步动作。下面介绍基本的过程。

初始化时,将启发式控制器的优先级 M_h 设置为一个不小的非负常量,并将选择概率 P_h 设置为不小的常量($0 < P_h < 1$),而SARSA控制器的 M_s 和 P_s 均设置为零。在求解初期,由于智能体的经验很少,$M_s < M_h$,所以一般会选择启发式控制器决定的动作,随着智能体经验的增加,M_h 随之增加。当 $M_s = M_h$ 时,就转入了以激励学习为主、启发式为辅的阶段。由上文所知,由于SARSA采取的是随

机策略,当 $P_s < P_h$ 时,意味着总控制器认为启发式信息是比较重要的,这样,总控制器将选择启发式控制器的决定;反之,当 $P_s > P_h$,选择 SARSA 控制器的决定。

3 实例分析

实验采用随机生成的迷宫模型,见图 1,其中 S 代表出发点(左下角),G 代表目的地(右下角),黑色为不可通行区域,白色为可通行区域。实验要求智能体从点 S 出发在尽可能短的时间内到达点 G,智能体在行进的过程中,有 4 个方向可以选择,即:上、下、左、右。智能体在执行每一步动作后,观察状态 s,得到瞬时奖赏值 r,其中到达点 G 时,$r = 1000$;经过可通行区域时,$r = -0.02$;碰到边界和不可通行区域时,$r = -5$。如果到达目标或者每次尝试超过规定步数时,开始新一次的尝试。文中设定的每次尝试的步数最多为 5000 次,取连续成功 100 次的数据。

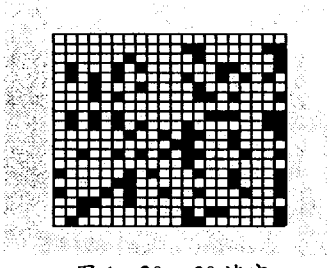

图 1　20 × 20 迷宫

根据实验环境的特点,启发式函数 $h(x)$ 采用最简单、最直观的距离启发式。智能体每走一步,就计算当前位置和目标的位置的距离,距离短将优先考虑。实验共采用随机生成的 10 个迷宫模型进行比较,对同一模型用启发式搜索、SARSA 和启发式 SARSA 分别进行实验(见图 2、图 3)。

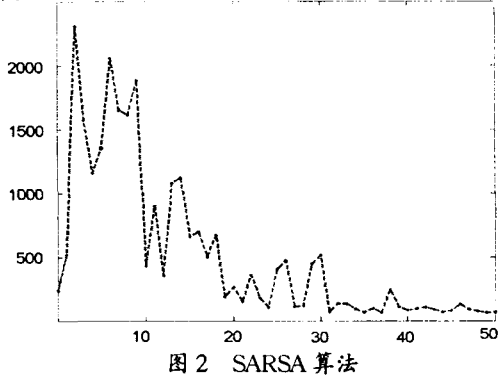

图 2　SARSA 算法

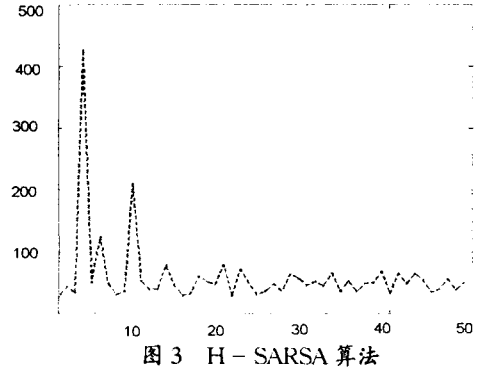

图 3　H - SARSA 算法

(注:图 2、3 的横坐标为收敛到最优所需的次数,纵坐标为成功时到达目标的步数)

通过实验可知,由于启发式函数设置的较为简单,而且为了节约空间,没有采用堆栈、回朔等办法,它的性能很不稳定,在较为简单的迷宫中,它能很快找到最优解,而稍复杂一点的迷宫中,它很可能陷入死循环,得不到解;对于 SARSA 学习进行的搜索,性能比较稳定,但是平均要失败 7 到 8 次才能到达终点;而对于基于启发式的 SARSA 而言,实验取得了较为满意的成绩,没有任何失败就可以到达终点。图 2 采用标准 SARSA 算法,可以看到智能体大概需要 30 次以上才能找到最优,且每次搜索寻找的步数也较多;而图 3 采用的是基于启发式搜索的 SARSA(H - SARSA),只需 10 次就能达到最优,且每次搜索寻找的步数很少。在简单的迷宫中,如果单纯的启发式能得到最优解,H - SARSA 也能很快得到最优解,而对于复杂的迷宫状态,它同样能较快地找到最优解。但是,要注意的是参数的设置,当用启发式搜索不能得到最优解时,可以适当降低 M 和 P 值,减少启发式信息带来的负面干扰。实验证明,P 取 0.3 时,效果最好。

4 结　论

结合了启发式搜索的激励学习算法在实验中证明是可行的,但是还有一些问题需要改进。比如,实验中参数的调整是通过人为调整的,这样比较麻烦而且很可能不精确,如果能加以改进,让智能体自适应调整将会更好。另外,采用总控制器和子控制器的分层结构可以使子控制器的增加变的容易,但是如何协调子控制器以取得更好的效果也是一个难题。对这些问题的改进将是下一步研究的重点。

参考文献:

[1] 王文杰.人工智能的原理与应用[M].北京:人民邮电出版社,2003.

[2] Crites R H, Barto A G. Elevator group control using multiple reinforcement learning agents[J]. Machine Learning, 1998, 33(2):235 - 262.

[3] Moore. A Variable resolution dynamic programming: Efficiently learning action maps in the real valued spaces[A]. In Proceedings of the 8th International Machine Learning [C]. Williamstown, Massachusetts, USA: [s. n.], 1991. 333 - 337.

[4] Rummery G, Niranjan M. On - line Q - learning using connetionist systems [R]. Technical Report, CUED/F - INFENG/TR166, Engineering Department, Cambridge University, 1994.

[5] Watkin C J, Dayan P. Q - Learning[J]. Machine Learning, 1992(8):279 - 292.

[6] Singh S, Sutton R S. Reinforcement learning with replacing eligibility traces[J]. Machine Learning, 1996(22):123 - 158.