重庆理工大学学报(自然科学)
Journal of Chongqing University of Technology(Natural Science)
2022 年 第 36 卷 第 12 期
Vol. 36 No. 12 2022
doi: 10.3969/j.issn.1674-8425(z).2022.12.016

“机器博弈”专栏(专栏主编:张小川 重庆理工大学 教授)

Q 学习实现亚马逊棋评估函数自调参

邱虹坤[1],王浩宇[1],王亚杰[2]

(1. 沈阳航空航天大学 计算机学院,沈阳 110136;
2. 沈阳航空航天大学 工程训练中心,沈阳 110136)

摘 要:在亚马逊棋评估函数模型中进行参数调试,主要由人工依靠经验反复实验来实现,效率较低且无法保证精度。针对人工调参效率低下、精确度不足的问题,可借助机器学习的方法来弥补。采用强化学习中 Q 学习的思路,构造一种具有自学习能力的网络结构,利用计算机自身反复模拟对局与迭代,实现评估函数调参工作的自动化。实验结果表明:当训练达 10 000 次时,模型各结点 Q 值会趋于收敛,说明此时程序可以做出稳定合理的调参操作;在博弈实战中,模型调参后的程序也表现出了较强的棋力。

关 键 词:计算机博弈;亚马逊棋;强化学习;Q 学习;评估函数
中图分类号:TP311 **文献标识码**:A **文章编号**:1674-8425(2022)12-0136-06

0 引言

计算机博弈是人工智能领域的重要研究方向之一。近年来,计算机博弈的快速发展引起了国内外专家学者的高度关注。亚马逊棋正是一种新兴的计算机博弈棋种,它是一种两人完全信息动态博弈游戏。棋子的移动方式类似于国际象棋中皇后的走法,同时具有很强的领土控制特色[1-3]。

评估函数是亚马逊棋博弈程序中最为核心的部分之一,其作用是评估当前棋局局面的优劣性并为其打分。该分数作为后续博弈树搜索等一系列操作的基础,评估函数是否合理将直接影响到整个程序的棋力。对于亚马逊棋评估函数中的一些关键参数,目前主要通过人工方法,根据经验反复进行实验调参后才能确定,这种方法效率较低且精确度无法保证。一些学者的研究表明,通过引入强化学习的方式,可在一定程度上提高调参的效率和精确性[4-5]。

本文采用强化学习中 Q 学习的思路,设计了一种能实现评估函数自主调参的网络模型。相较于传统调参方法,该模型节省了大量的人力与时间,并且通过自主训练,对参数的调节也更加精确

收稿日期:2022-09-02
基金项目:辽宁省兴辽英才计划项目(XLYC1906003);辽宁省教育厅科学研究项目(JYT2020038);沈阳航空航天大学大学生创新创业训练计划项目(202010143008)
作者简介:邱虹坤,男,硕士,副教授,主要从事机器博弈研究,E-mail:qiuhongkun@sau.edu.cn;通讯作者 王浩宇,男,主要从事机器博弈研究,E-mail:1049854191@qq.com。

本文引用格式:邱虹坤,王浩宇,王亚杰. Q 学习实现亚马逊棋评估函数自调参[J]. 重庆理工大学学报(自然科学),2022,36(12):136-141.
Citation format: QIU Hongkun, WANG Haoyu, WANG Yajie. Parameter self-adjustment of Amazon Chess evaluation function through Q-Learning[J]. Journal of Chongqing University of Technology(Natural Science),2022,36(12):136-141.

合理。

1 强化学习和 Q 学习

1.1 强化学习

在监督学习方式中,数据集的精确性很大程度上决定了训练结果的优劣性。然而在某些情况下,要生成大量准确合理的教师数据是十分困难的,如对棋盘博弈游戏或电子游戏的训练等。相比之下,强化学习更能胜任这些工作。强化学习是指在一连串行动的最后进行评价的学习方式。它不依赖于海量数据集的指导,而是通过不断与环境交互,总结经验并对行为的好坏进行评价打分,基于这些评价来推进学习[6-8]。该评价值被称之为奖赏(reward)。当强化学习到达某结果并获得奖赏时,与之关联的每一个行动都将被分配一定奖赏,以此来更新学习网络。

1.2 Q 学习(Q-learning)

Q 学习是最重要的强化学习算法之一,在该学习框架中,定义学习对象为 Q 值(Q-value),代表选择下一个行动的指标数值集合,通常建立一个 Q 表格来存储。Q 学习的最大特点在于将动态规划与蒙特卡洛方法结合,用于计算强化学习的整个马尔科夫奖励过程,得出最优策略[9-10]。其中马尔科夫过程又称马尔科夫链,指一系列不依赖于历史,仅根据当前来决定未来的状态集合,组成的无记忆的随机过程。而奖励过程则是在其基础上引入了奖励系数与衰减系数,因而可计算出某奖励链条上的收获(Return)[11]。

在训练初期,Q 值被随机分配,此时相当于随机选择行动。在更新 Q 值时可以遵循以下规则:在行动获得奖赏时,增加与其相关行动的 Q 值;在没有获得奖赏时,利用与下一状态相关联行动的 Q 值来更新当前 Q 值[12]。如式(1)所示。

$$Q(S_t,a_t) = Q(S_t,a_t) + \alpha(r + \gamma \max Q(S_{t+1},a_{t+1}) - Q(S_t,a_t)) \quad (1)$$

其中: s、a 分别代表 t 时刻的状态以及选择的行动,r、γ、α 分别代表获得的奖励值、折扣率以及学习系数[13]。学习系数是用于调节学习速率的常数,一般取 0.1 左右;折扣率的引入用于避免产生状态的无限循环,一般取 0.9 左右[14]。

2 亚马逊棋评估函数设计

根据亚马逊棋的游戏规则,游戏一方的最终目的是用自己的棋子和障碍挡住对手的棋子,因此一般采取占领领土或阻塞对手路线的思路来设计评估功能[15]。亚马逊棋评估函数的设计主要是基于 kingmove 和 Queenmove 走法。Queen 走法是按照国际象棋中的皇后走法,可以向横竖斜对角 8 个方向走直线,只要路径上没有障碍就可以沿直线一直走;king 走法则是按照国际象棋中的国王走法,即只能向 8 个方向走一格。设 t_1、t_2 为 territory 特征值,代表双方对空闲区域的控制能力的评估值;c_1、c_2 为位置 position 特征值,用于反映双方对空格控制权的差值特征。a、b、c 参数是根据棋局进行程度 w 而不断变换的权重,用于动态控制在不同棋局进度下各个特征值对结果的影响[16]。此外,为避免出现区域浪费情况(某空区域被围堵,不可再被使用),定义参数 S 用于判断当前步法是否会产生该种区域的出现(称为缺陷区域)。需要注意,这里的"S"与式(1)中的"S"含义不同。式(1)中"S"代表强化学习模型中 t 时刻的状态;而这里的"S"是为避免出现区域浪费情况而在亚马逊棋评估函数中引入的参数。综上,可以得出评估函数的最终表达式,如式(2)所示。

$$Value = a \times t_1 + b \times t_2/2 + c \times ((c_1 + c_2)/2) + S \quad (2)$$

a、b、c 3 个权重系数的确定对最终估值精度有至关重要的作用。一个根据经验确定的权重与进度系数 w 的映射关系如式(3)所示[17]。其中,X 值与 Y 值是决定评估准确性较为关键的参数,也是该评估函数中主要调参对象。

$$\begin{cases} a = X/(w + X) \\ b = w/(w + Y) \\ c = (1 - (a + b))/2 \end{cases} \quad (3)$$

此外,虽然通过计算当前局面的领土特征值和棋子的灵活度等参数可以得到一个较为精确的评估价值,但有时仍会存在偏差,可以采用一种更为简单直观的方式来校准一下最终估值。通过"自对弈"的思路进行矫正,即以当前局面为基础,电脑控制双方随机走棋(在一个辅助棋盘中),在

合理时间内进行大量对局,从而计算出胜率。随机下棋虽然无规律性,但是模拟速度快。当进行大量模拟时也能一定程度上反映当前局面的优劣性。将模拟出的胜率与评估函数结果相结合即可达到调整效果。

3 Q 学习调参模型

3.1 调参模型设计

本文主要的调参对象即为式(3)中的 X 值与 Y 值,可以采用迭代的思路实现自动调参,简单来说就是对 X 参数或 Y 参数分别进行增大或减小,。首先定义 2 个数值 C_1、C_2,分别代表对 X、Y 进行加或减操作时的变化值。以 X、Y 的具体数值代表当前状态,而变更数值的方式则称为“行动”。因此,结合一般手动调参的方法,对于每种情况都有 6 种后续可能的行动,如表 1 所示。表中前 4 种行为对应的就是“增大 X”“减小 X”“增大 Y”以及“减小 Y”。而 5、6 号行动的引入分别对 C_2 进行加操作以及对 C_1 进行减操作,目的是避免发生多次行动后的参数恰好未发生变化,导致调参陷入死循环。对此,建立一个三层六叉树作为整体学习训练框架,共 43 个结点,如图 1 所示。

表 1 后续行动方式

编号	行动
1	$X = X + C_1;\ Y = Y$
2	$X = X - C_2;\ Y = Y$
3	$X = X;\ Y = Y + C_1$
4	$X = X;\ Y = Y - C_2$
5	$X = X + C_1;\ Y = Y + C_2$
6	$X = X - C_1;\ Y = Y - C_2$

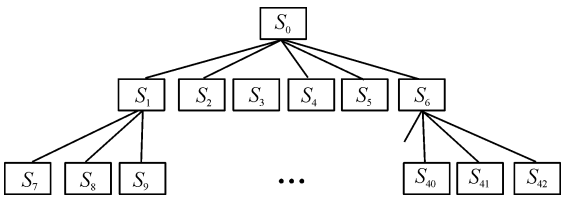

图 1 三层训练网络结点关系图

模型利用对局的胜负情况来确定是否进行奖赏。首先记录下初始的参数值: X_0、Y_0,而后经过一系列行动选择后获得新的参数值: X_1、Y_1。将新旧参数带入博弈程序中调用,并进行数局对弈[18]。若新参数的胜率大于旧参数,则可以认为这是一种较好的调参策略,并对该行动路径上所有结点的 Q 值进行奖赏操作。之后再重新选择行动,获得新的 X_1、Y_1 并进行奖赏评估。反复执行此过程后,最终会得到一条较为稳定的,Q 值相对较高的行动路线[19]。按照该行动路线更新评估函数参数,即代表完成了一次调参操作。之后初始化整个训练模型,以更新后的评估参数为基础,进行新一轮训练。如此反复以达到自调参的目的。整个训练流程如图 2 所示。

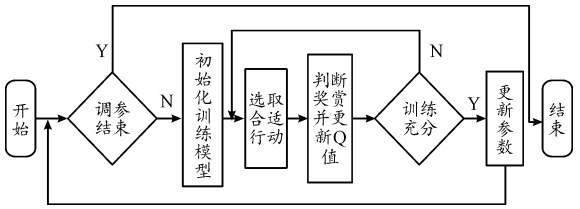

图 2 训练流程框图

3.2 调参模型实现

调参模型是在博弈程序的基础上构造而成。首先定义节点型结构体 node,内容包括每个结点的 Q 值,对 X 参数和 Y 参数所进行的操作,以及其 6 个孩子结点,可利用一个数组存储。定义全局变量 Weight,用于存放参数值; 定义一个整形数组 path,用于存放选择过的行动路径。

程序主要包括 3 个功能函数: 为模型初始化、行动判断、Q 值更新。模型初始化时要求对所有结点信息进行初始化,包括 Q 值的随机分配、行动定义以及孩子结点的链接等。行动选择的基本思路是优先选择 Q 值最大的行动,但可能会出现某一高 Q 值行动被反复选择,同时其他行动无法得到训练的情况。该问题可以通过 ε-贪心法(ε-greedy)的思路解决。首先确定一个 ε 值并生成一个 0 ~ 1 的随机数,若该随机数小于 ε,则忽略 Q 值进行随机行动选择; 若大于 ε,则选取 Qmax 行动[20]。

Q 值更新主要基于式(1)进行计算,将学习系数设为 0.1,折扣率设为 0.9。首先执行新旧参数的自对弈功能,通过对弈结果判断是否奖赏。如果能获得奖赏,则将奖赏按比例加到 Q 值上; 若不

能,则从下一状态所能选择的行动中对应的 Q 值,将最大值按比例加到 Q 值上。由于叶子结点后无其他行动可选择,因此对于模型第二层和第三层的 Q 值更新需要分别采取上述 2 种计算方法。

在实际调参过程中,Q 值会影响程序调参的方法,从而得出不同的参数 X 和 Y。而 X 与 Y 会进一步影响整个评估函数中不同参数的权重比,最终得出评估函数的计算最终结果 Value。

4 实验测试

首先初始化训练参数。设评估函数模型中待调整的参数 X、Y 初值分别为 5 和 20。对于训练模型的每次调参操作中,训练次数为 10 000 次,即模型在每次调参前需要更新 10 000 次网络 Q 值。以黑方棋子作为调参测试对象。规定每次自对弈的局数为 5 局,即若黑方获胜 2 局以上即可获得奖赏。为了加快调参速度,限制了较浅的搜索深度以及较短的搜索时间。在对弈时,除了双方程序评估函数的参数不同之外,其内部算法和逻辑结构等部分完全相同。

对模型进行第一次调参训练,从中随机抽取并跟踪了 6 个结点的 Q 值变换过程,分别为网络第二层的 3 号、5 号结点以及第三层的结点 20 号、22 号、33 号和 36 号结点,结果如图 3、图 4 所示。可见,当训练次数为 1 000 时,Q 值仍处于动态变化之中,在后期开始出现收敛的趋势。而当训练次数达 10 000 次时,各结点的 Q 值已处在一个相对稳定的区间内波动,从而可以认为此时的网络已经训练的较为充分。

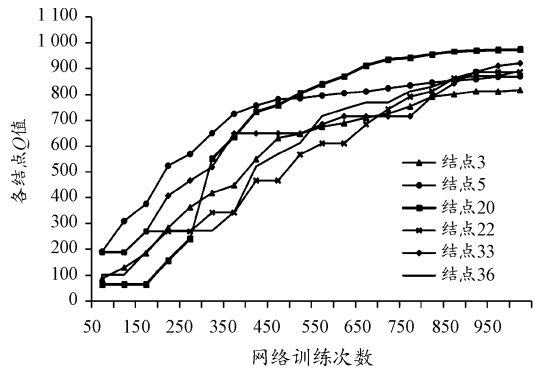

图 3 训练 1 000 次时节点 Q 值变化曲线

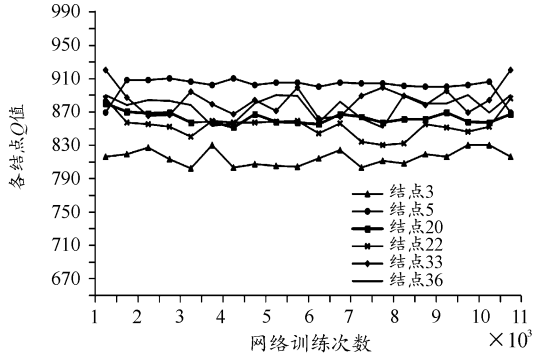

图 4 训练 10 000 次时节点 Q 值变化曲线

因此,针对本次训练的 6 个结点而言,得出的行动路线是 5 号结点至 33 号结点,最终的调参步骤应是先进行 5 号结点操作($X = X + C_1$、$Y = Y + C_2$) ,之后进行 33 号结点操作($X = X$、$Y = Y + C_1$) 。重复上述调参步骤,进行数次调参后最终参数 X 与 Y 的变化情况分别如图 5、图 6 所示。可见,在训练过程中 X、Y 的值逐渐收敛并趋于稳定。训练结束后,X 值由最初的 5 被调整为 3,而 Y 值由最初的 20 被调整为 19. 2。

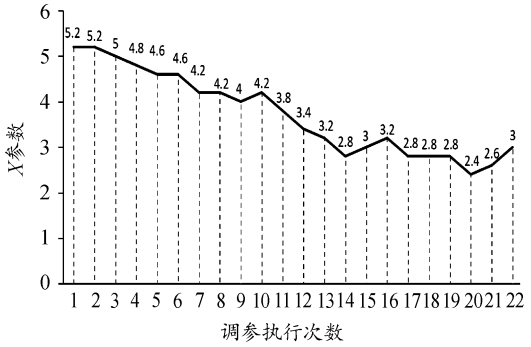

图 5 参数 X 调整情况

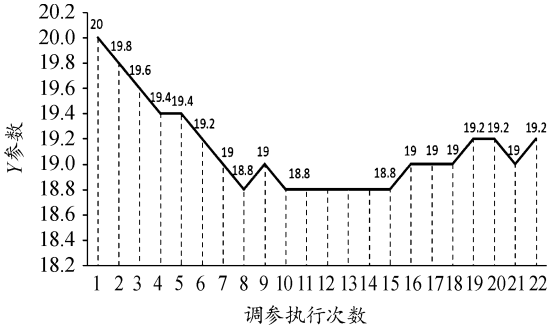

图 6 参数 Y 调整情况

理论上,在单次调参操作中进行更多的网络训练(即 Q 值更新) ,程序会选择更加合理的调参

行动。然而,每次的调参操作对整体评估函数参数的影响是有限的,且每次执行调参后该网络模型都会被重置初始化。因此,在单次调参中对网络进行过多的训练,性价比显得不高。应结合具体的硬件设备和程序本身来确定一个较为合适的单次调参训练次数,尽可能多地完成调参操作,而避免花费大量的时间在仅仅一次的调参操作内。这样,调参的效果会更加明显,调参的整体效率也会更高。

为验证程序整体棋力水平,该程序参加了中国计算机博弈锦标赛。在比赛过程中,相较其他程序而言,其最大特点是对棋盘空位的利用率很高,能更有效地利用自己目前所占有的区域。在与强队进行比赛时,双方棋力相差不大,我方凭借对棋盘空位的充分利用,最终靠细微优势打败对手,如图 7 所示,黑棋为我方棋子。最终该程序取得了季军的成绩,说明该程序棋力优于大部分传统博弈程序,同时调参的效果也得到了体现。

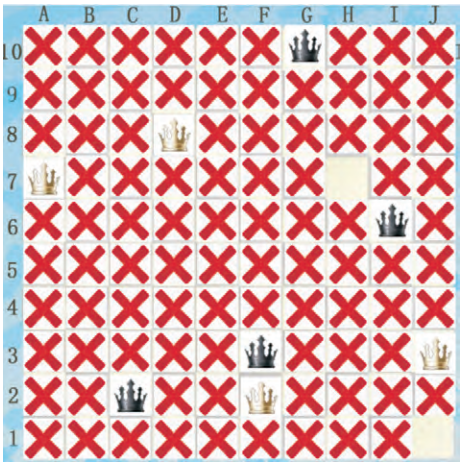

图 7 对弈终局局面

5 结论

结合 Q 学习的思想,构建了一个亚马逊棋评估函数自调参强化学习模型。克服了传统人工调参方式耗时耗力的缺点,实现了调试的自动化,同时对棋力的提升起到了一定效果。该思路不仅适用于亚马逊棋的调参工作,对于类似的完全信息博弈类棋种依然适用。

参考文献:

[1] 王亚杰,邱虹坤,吴燕燕,等. 计算机博弈的研究与发展[J]. 智能系统学报,2016,11(6):788 – 798.

[2] 吕艳辉,宫瑞敏. 计算机博弈中估值算法与博弈训练的研究[J]. 计算机工程,2012,38(11):163 – 166.

[3] 王亚杰,王晓岩,邱虹坤,等. 建设棋牌谱标准 构建计算机博弈竞赛持续发展新生态[J]. 实验技术与管理,2020,37(2):19 – 23.

[4] 陈圣磊,谷瑞军,陈耿,等. 基于 TD(λ) 的自然梯度强化学习算法[J]. 计算机科学,2010,37(12):186 – 189.

[5] 陈兴国,俞扬. 强化学习及其在电脑围棋中的应用[J]. 自动化学报,2016,42(5):685 – 695.

[6] 杜康豪,宋睿卓,魏庆来. 强化学习在机器博弈上的应用综述[J]. 控制工程,2021,28(10):1998 – 2004.

[7] 毛健,赵红东,姚婧婧. 人工神经网络的发展及应用[J]. 电子设计工程,2011,19(24):62 – 65.

[8] 刘全,翟建伟,章宗长,等. 深度强化学习综述[J]. 计算机学报,2018,41(1):1 – 27.

[9] 赵冬斌,邵坤,朱圆恒,等. 深度强化学习综述: 兼论计算机围棋的发展[J]. 控制理论与应用,2016,33(6):701 – 717.

[10] 卢海峰,顾春华,罗飞,等. 基于深度强化学习的移动边缘计算任务卸载研究[J]. 计算机研究与发展,2020,57(7):1539 – 1554.

[11] WEI Q L, LIU D R, SHI G. A novel dual iterative Q-learning method for optimal battery management in smart residential environments[J]. IEEE Transactions on Industrial Electronics,,2014,62(4):250942518.

[12] TANG R, YUAN H. An error-sensitive Q-learning approach for robot navigation[C]//2015 34th Chinese Control Conference (CCC). New York: IEEE, 2015: 5835 – 5840.

[13] 司彦娜,普杰信,臧绍飞. 基于残差梯度法的神经网络 Q 学习算法[J]. 计算机工程与应用,2020,56(18):137 – 142.

[14] 小高知宏. 机器学习与深度学习: 通过 C 语言模拟[M]. 申富饶,于傲,译. 北京: 机械工业出版社,2018.

[15] 郭琴琴,李淑琴,包华. 亚马逊棋机器博弈系统中评估函数的研究[J]. 计算机工程与应用,2012,48(34):50 – 54.

[16] JENS L. An evaluation function for the game of amazons[J]. Theoretical Computer Science,2005,349: 232 – 234.

[17] 陈萱华,杨玲. 亚马逊棋中评估函数的研究[J]. 电脑知识与技术,2019,15(8):224 -226.
[18] SILVER D,HUANG A,MADDISON C J,et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature,2016,529(7587):484 -489.
[19] SALLAB A,ABDOU M,PEROT E,et al. Deep reinforcement learning framework for autonomous driving[J]. Electronic Imaging,2017,19:70 -76.
[20] SILVER D,HUBERT T,SCHRITTWIESER J,et al. A general reinforcement learning algorithm that masters chess,shogi,and go throughself-play[J]. Science,2018;362(6419):1140 -1144.

Parameter self-adjustment of Amazon Chess evaluation function through Q-Learning

QIU Hongkun[1], WANG Haoyu[1], WANG Yajie[2]

(1. School of Computer Science, Shenyang Aerospace University, Shenyang 110136, China;
2. Engineering Training Center of DUT, Shenyang Aerospace University, Shenyang 110136, China)

Abstract: Parameter adjustment of Amazon Chess evaluation function mainly relies on manually repetitive experience, which is inefficient and the accuracy is hard to be guaranteed. Aiming at the problems of low efficiency and insufficient accuracy of manual parameter adjustment, this paper proposes a machine learning to solve them. By using Q-learning in reinforcement learning, a network structure with self-learning abilities is constructed, and parameter self-adjustment of the evaluation function is achieved through repetitive self-simulation of games and iteration by the computer. The experiment result shows that, after the training is done for at least 10 000 times, the Q value of each node of the model tends to converge, which indicates that the program can make stable and reasonable parameter adjustment. In real matches, the program of the adjusted model also behaves well.

Key words: computer-based game; Amazon Chess; reinforcement learning; Q-learning; evaluation function

(责任编辑 王 欢)