DOI:10.3969/j.issn.1673-4785.201110005
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20120518.0845.002.html

采用时间差分算法的九路围棋机器博弈系统

张小川,唐艳,梁宁宁
(重庆理工大学 计算机科学与工程学院,重庆 400054)

摘 要:围棋机器博弈是机器博弈中重要的分支之一,其庞大的博弈空间给机器博弈研究者带来了巨大挑战.目前围棋机器博弈多采用静态估值搜索与蒙特卡洛树搜索,故将时间差分算法引入至九路围棋机器博弈系统中,提出基于时间差分算法的围棋机器博弈系统模型,该博弈系统具有一定的自学习能力,能在不断的对弈中逐步提高博弈能力.通过与采用α-β搜索算法的博弈系统进行实际对弈,证明了该方法的可行性.

关键词:机器博弈;九路围棋;围棋机器博弈;时间差分算法
中图分类号:TP31 **文献标志码**:A **文章编号**:1673-4785(2012)03-0278-05

A 9×9 Go computer game system using temporal difference

ZHANG Xiaochuan, TANG Yan, LIANG Ningning
(College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China)

Abstract: Computer Go is an important branch of computer games and presents great challenges to computer game researchers due to its need for huge game space. Presently, the static evaluation method and the Monte-Carlo tree search method are widely used in Go computer games. In this paper, a temporal difference algorithm was introduced to the 9×9 Go computer game system which gave it self-learning capability, thereby improving the game levels as a result of the continuous training. Through playing chess with a system which adopts an α-β algorithm, the new method was proven to be effective.

Keywords: computer game; 9×9 Go; Go computer game; temporal difference

近年来人工智能是信息科学中重要的热点研究领域之一,其相关算法、技术及研究成果正被广泛运用于各行业,如军事、心理学、智能机器、商业智能等.机器博弈是人工智能研究的重要分支,而围棋机器博弈是机器博弈的热点问题之一,其庞大的搜索空间和较高的复杂度,使其在机器博弈中有着重要的研究价值.

目前,围棋机器博弈中常采用的博弈算法有α-β剪枝搜索算法[1]、模式匹配[2-3]和UCT算法[4]等.围棋机器博弈相对于六子棋、象棋等其他棋类博弈拥有更大的搜索空间和更高的复杂度,当采用α-β等传统搜索算法时,会在时间有限情况下无法搜索到目标解.因此,本文尝试将时间差分法引入至围棋机器博弈,将博弈系统看成一个具有自我学习能力的围棋人工生命体或围棋智能体,它能在不断的博弈过程中提高自己的博弈能力.借助计算机C语言,实现了该围棋机器博弈系统,并且通过博弈实战验证了该方法的有效性和可行性.

1 时间差分算法

1.1 强化学习

强化学习(reinforcement learning, RL)较其他常用的机器学习算法,如神经网络、决策树等,在博弈系统中有着独特优势.该方法通过不断的试探与环境进行交互,根据试探所得到的反馈来决定下一动作选取,不同于传统的监督学习.监督学习需要一个教师信号来告诉智能体怎样选取动作,并给出好坏程度的评价标准,而强化学习则是通过环境反馈来

收稿日期:2011-10-17. 网络出版日期:2012-05-18.
基金项目:重庆市教委科研项目(KJ120824);重庆市自然科学基金资助项目(2007BB2415).
通信作者:张小川. E-mail: cqpczxc@163.com.

评价采取某动作的好坏[5-7]. 图 1 简单描述了强化学习中智能体与环境的交互过程(其中 s 为智能体当前所处的环境状态).

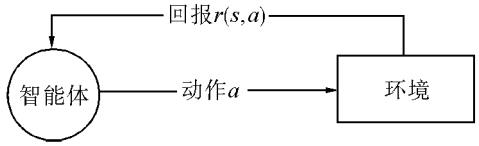


图 1　强化学习中智能体与环境交互过程

Fig. 1　The agent and the environment interaction process in reinforcement learning

如果将强化学习应用到围棋机器博弈中,博弈程序变成具有一定智能的决策者,而围棋棋盘就被看作博弈环境. 当博弈双方产生新着法时,围棋棋盘的状态就发生了改变,博弈环境的状态也随着发生转移. 同时在博弈进程中,从博弈开始到博弈结束,其整个过程包含系列的博弈着法,即博弈着法的集合;因此利用强化学习解决围棋博弈问题的核心,就是要建立一种合适的内部奖励机制,使得博弈程序或围棋人工生命体能执行最大化内部奖励的局部动作,从而学会发现一个最佳的着法序列,并提高博弈水平.

1.2　时间差分算法

时间差分算法(temporal difference)是强化学习的一种重要算法[7],其利用探索所得到的下一状态的价值和奖励来更新当前状态的价值[8]. 本文经过研究分析,构造了博弈状态转移特征方法,利用该方法获得的特征信息(特别是激励性信息)反馈于当前博弈状态,并更新当前博弈状态,引导博弈系统的价值取向,这就是本文引入时间差分算法的机器博弈的基本思路.

在实际应用中,通常采用成对的状态-动作值 $Q(s_t,a_t)$ 来表示当处于状态 s_t 时执行动作 a_t 的价值. 在简单的确定的情况下,任意一对状态-动作只有 1 个奖励和可能的下一状态,根据 Bellman 公式,可得如式(1)的简化公式[9]:

$$Q(s_t,a_t) = r_{t+1} + \gamma \max_{a_{t+1}} Q(s_{t+1},a_{t+1}). \quad (1)$$

式中:$\max\limits_{a_{t+1}}(s_{t+1},a_{t+1})$ 表示充分利用已有发现,选择具有最高价值的动作. 当处于探索阶段时,若处于状态 s_t,则随机选取一个动作 a_t,返回一个奖励 r_{t+1},并将状态转移至 s_{t+1}. 此时,前一动作的价值更新为

$$\hat{Q}(s_t,a_t) \leftarrow r_{t+1} + \gamma \max_{a_{t+1}} \hat{Q}(s_{t+1},a_{t+1}). \quad (2)$$

由此可以看出 $\hat{Q}(s_{t+1},a_{t+1})$ 是更新后的值,具有更高的正确概率. 将式(2)引入,以减小当前 Q 值与一个时间步骤之后的估计值之间的误差,则有式(3):

$$\hat{Q}(s_t,a_t) \leftarrow \hat{Q}(s_t,a_t) + \eta(r_{t+1} + \gamma \max_{a_{t+1}} \hat{Q}(s_{t+1},a_{t+1}) - Q(s_t,a_t)). \quad (3)$$

式中:η 为更新因子,随时间的增加逐渐减小;γ 为折扣率,$0 \leqslant \gamma < 1$,保证返回的奖励为有限的.

对于动作的选取,在知识量少的初期,可以在所有动作中随机选取,可看作“探索”. 但也不希望一直探索下去,故探索到一定时,需利用当前所学知识. 为此采用一个温度变量 T 来实现从探索到利用知识的转移,下面给出加入温度变量时选择动作 a 的概率[10]:

$$P(a \mid s) = \frac{\exp[Q(s,a)/T]}{\sum_{b=1}^{A} \exp[Q(s,b)/T]}. \quad (4)$$

当 T 很大时,所有概率趋近于相等,此时进行随机探索;当 T 很小时,价值更大的动作被选取的可能性较大,则实现对知识的利用. 所以在学习的过程中以一个较大的 T 值开始,不断地缩小 T 值,完成探索直至利用知识.

2　基于时间差分算法的围棋机器博弈模型

当求解问题的状态空间较大时,会使强化学习算法的收敛效率降低,这就要求增加实验次数,但降低了算法的实时性[11]. 而在围棋机器博弈中,若搜索超时则直接判负. 并且当处于中局时,棋盘状态复杂度增加,若把每个可下点看作一个动作,则算法的状态与动作数量大幅度增长. 故需采用其他策略减少问题状态空间,以增强算法的实时性. 为此,采用将静态估值与时间差分算法相结合的策略,在产生可下节点时,选取静态估值较大的点,再在此点上利用时间差分算法完成动作的选取.

2.1　系统状态

在博弈过程中,围棋棋盘状态作为环境因素直接影响博弈智能体作出的决策,如开局时摆棋形、博弈过程中己方受威胁棋子、对方受威胁棋子等. 本文选取环境因素中对博弈智能体的决策影响较大的因素作为系统问题状态. 该状态集形式化描述如式(5):

$$S = \{S_n,S_e,S_l,O_n,O_e,O_l\}. \quad (5)$$

式中:S_n 为当前棋盘上己方棋子总数,S_e 为当前棋盘上己方眼总数,S_l 为当前棋盘上己方气总数,O_n 为当前棋盘上对方棋子总数,O_e 为当前棋盘上对方眼总数,O_l 为当前棋盘上对方气总数. 其中,S_n 与 O_n 直接关系到当前博弈双方对弈的局势;S_e 与 O_e 直接关系到某串棋是否为活棋,如当某串棋有 2 个眼,则被提掉的可能性减小至 0;S_l 与 O_l 则直接关

系到棋子被提的可能性和地盘的占有率.

2.2 系统动作

围棋每走一步都有相应的说法,即术语,而常用的围棋术语有很多,如“拆”、“飞”、“长”、“立”、“尖”、“扳”、“接”、“断”、“挖”、“夹”、“托”、“虎”和“刺”等[12].若将每一种下法作为一种动作,则系统动作数量会过大而使算法失去实用性.这需要将术语归类,也就是划分基本动作.下面以“扳”和“挖”为例说明其归为哪一基本动作,如图2所示,未标号棋子为已下棋子,标号棋子为欲下棋子.

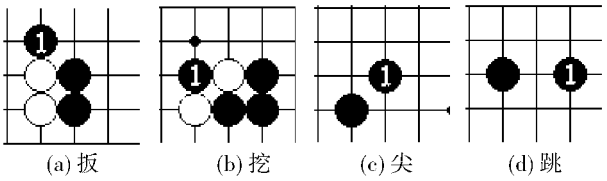


图2 围棋的一些着法

Fig. 2 Some actions in Go

由图2可知,“扳”和“挖”均可看作在己方棋子的“尖”或“跳”位置上下棋,若选取“扳”和“挖”中离棋子1位置最近的己方棋子则为“尖”.采用此归类方法,选取如下四大类下法作为基本动作:“拆”、“飞”、“尖”和“长”,如图3所示.

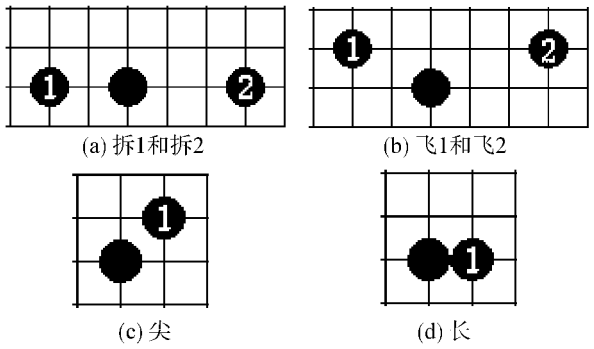


图3 四大基本动作

Fig. 3 4 Basic actions

由此可得到动作的形式化描述集 A,具体形式如下:

$A = \{\langle \text{ActionType}\rangle, \langle \text{ActionDirection}\rangle\}$;

ActionType = {extend1, entend2, knimove1, knimove2, diamove, nobi};

ActionDirection = {up, down, left, right, topright, botright, topleft, botleft}.

式中:动作集 A 由动作类型和方向参数2个参数组成.动作类型有6种,分别为“拆1”、“拆2”、“飞1”、“飞2”、“尖”和“长”.“拆1”、“拆2”、“尖”和“长”的方向参数为上下左右4个,“飞1”和“飞2”的方向参数为上下左右和右上、右下、左上、左下8个.6种动作组合一起,共32个动作.将时间差分算法应用在围棋机器博弈中,则此时需解决的问题转化为求合适的<类型,方向>动作.

2.3 动作奖励

当尝试动作 a 时,系统会获得一个奖励 r_a,并且在围棋机器博弈中这样的奖励是确定的.在实际的博弈过程中,奖励跟下棋后棋盘位置的静态值、己方棋子总数与对方棋子总数、是否吃子与被吃、气微薄的数目等信息有关.例如,当落下某棋子时,使得某串棋的气数减少(甚至为1),这样很有可能在对方下一手棋或后几手的时候提掉整个串,这样的下子动作将会得到较少的奖励(甚至为负).基于这样的情况,下面给出动作奖励规则:

$$r_a = S_v + S_n + L_n + S_1. \tag{6}$$

式中:S_v 为棋盘棋子位置的静态分值,S_n 为己方棋子总数与对方棋子总数的差值,L_n 为吃子与被吃子数的差值,S_1 为对方气为1的棋子数目与己方气为1的棋子数目的差值.

3 实 验

3.1 生成训练集

当 $Q(s_t, a_t)$ 值(即状态-动作值)很大时,用表格等手段存储,则表格的尺寸会非常大,这使得搜索空间也增大.为此,在基于时间差分算法的围棋机器博弈系统中,采用人工神经网络作为回归器,此时以 s_t、a_t 作为网路输入,$Q(s_t, a_t)$ 值为网络输出,如图4所示.

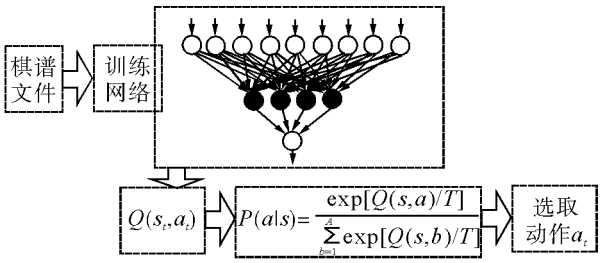


图4 采用神经网络的时间差分算法

Fig. 4 The flow chart of the application of temporal difference using neural networks

由于人工神经网络为监督学习方法,因此需要训练集TS.故本文首先将棋谱文件导入至博弈系统中,按照棋谱文件下棋,根据式(3)计算 $Q(s_t, a_t)$ 值,式(3)中需要用到的奖励 r_a 则由式(6)得到,再将 s_t、a_t、$Q(s_t, a_t)$ 存储至系统中得到样本集TS(表1给出样本集TS中10个训练样本).其中折扣率 γ 取0.5,η 取0.4,η 随时间逐渐减小,每次减小0.0012.需注意的是,由于围棋博弈空间巨大,故训练时需要相当数量的样本才能达到训练效果,本文

选取的样本数为4 000.

表 1　TS 样本集中的 10 个样本

Table 1　10 samples of TS sample set

s_t	a_t	$Q(s_t,a_t)$
880682780	(0,14)	0.237 247
381395919	(0,2)	0.517 648
347699070	(10,11)	−0.083 999
402741079	(12,10)	−0.093 058
820219960	(12,12)	0.472 617
709852749	(12,11)	0.451 825
514998196	(6,5)	−0.387 498
105471686	(11,10)	0.519 132
1016084926	(9,11)	0.501 035
56907172	(13,3)	0.506 086

3.2　仿真 $Q(s_t,a_t)$ 值与选取动作

本文采用 BP 神经网络,网络输入层有 9 个节点,由系统状态集 $S=\{S_n,S_e,S_l,O_n,O_e,O_l\}$、表示动作位置的 x 和 y,以及用哈希值存储的整个棋盘表示 s 组成,隐藏层 4 个节点,输出层 1 个节点. 初始训练时网络权值随机赋值,学习率 α 取 0.5,学习精度 θ 取 0.000 1. BP 神经网络根据前向传播输出原理,利用误差反向传播修改权值和阈值. 在学习过程中,可将每一个或一定数量的棋谱文件视为学习的一个停顿. 训练好神经网络后,保留修改好的权值和阈值等参数.

训练结束后,就可进行对弈. 在博弈时将提取到的棋盘状态 s_t 和搜索到的所有合法动作 a_t,输入至 9×4×1 的 BP 神经网络中,得到 $Q(s_t,a_t)$ 值. 然后把当前所有合法动作 a_t 所对应的 $Q(s_t,a_t)$ 值都求出来,之后便采用式(4)的方法选取动作 a_t. 其中式(4)中温度变量 T 的初值为 500,在博弈过程中逐渐减小(每次减小 1),从而达到从知识的探索过渡到知识的利用. 当 T 值减小到一定程度时则实现知识利用,$P(a|s)$ 值大的动作更容易被选取到. 此时本文采用轮盘赌的方式,生成一个 $p(0<p<1)$,判断 p 值落在哪 2 个动作的 $P(a|s)$ 值之间,便可判断选取哪个动作 a_t.

3.3　实验结果

在实验初期,由于采用零知识学习,未给予任何其他相关辅助知识,如眼的识别判断、活棋的判断等;故此时该博弈系统并没有体现其优势,常走出坏招死招. 当加入知识判断时,系统的博弈能力明显提高. 并且通过实验发现,在单纯采用时间差分算法时,博弈智能体在博弈初期发挥较好,搜索时间短,能为后面棋局摆一个良好阵形. 但当进入至中局和终局时,进攻能力减弱,系统处于劣势.

将引入时间差分算法的 CQUTGO-2 与采用 α-β 算法的 CQUTGO-1 对弈 100 盘,其中 CQUTGO-2 执黑和执白各 50 盘,其对弈结果如图 5 所示. 由此可见在采用时间差分算法后,博弈系统的博弈能力较之前有所提高.

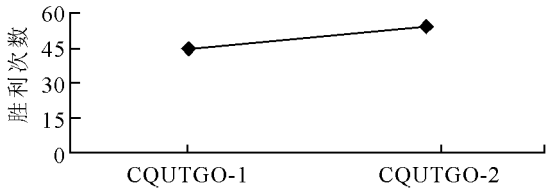


图 5　CQUTGO-1 与 CQUTGO-2 对弈的结果

Fig. 5　The game results of CQUTGO-1 and CQUTGO-2

3.4　结果分析

在基于人工神经网络的时间差分算法中,神经网络的各个方面均对算法在九路围棋机器博弈系统的应用效果产生影响,包括样本集、神经网络结构和训练次数.

1)样本集的选取. 在实际博弈过程中,同一系统状态下有多种动作可供选择. 采用棋谱文件导入至系统中,便于样本提取并可按不同对手选择不同的棋谱文件. 但棋谱文件中出现棋盘状态相同的次数较少,会降低样本集学习价值,影响学习效果. 有的学者在选取样本集时采用随机扩展方法,以产生在数量和质量上均可观的样本集[13].

2)神经网络结构. 采用神经网络仿真 $Q(s_t,a_t)$ 值时,网络输出则直接为相应的 s_t、a_t 的 $Q(s_t,a_t)$ 值. 故网络结构直接影响 $Q(s_t,a_t)$ 值,也就直接影响动作的选取和博弈的决策. 选取 9 个棋盘特征作为网络输入,但事实上这样并不能完全描述整个棋盘状态. 例如可将气为 1、气为 2 的棋子数作为棋盘特征时,当气为 1 时很可能被提掉,当气为 2 时,可以形成真眼.

3)训练次数. 在神经网络中,网络训练次数也直接关系到参数是否达到目标精度,直接影响学习效果.

4　结束语

本文将时间差分算法应用在机器博弈中,给出了包含系统状态、系统动作及动作奖励的博弈系统模型,并通过实验验证了该方法的有效性. 引入时间差分算法后的博弈系统是一个具有自学习能力的博弈智能体,能在不断的博弈过程中提高博弈水平. 由

于围棋博弈的复杂度较高,因此为了提高算法实时性,采用此类模型时将系统状态统计为6个状态因素向量,下棋动作划分为6类.这样便简化了系统状态和动作.虽然该方法能提高算法实时性,但其也存在不足,无法清晰划分动作和系统状态.而且系统状态和动作的划分直接影响人工神经网络结构,进而影响模拟结果.本文后期研究工作的方向是在保证算法实时性的前提下,如何划分系统的状态和动作.而现阶段围棋机器博弈大都采用蒙特卡洛算法,后期亦可考虑与其结合来提高算法的有效性.

参考文献:

[1]张聪品,刘春红,徐久成.博弈树启发式搜索的 α-β 剪枝技术研究[J].计算机工程与应用,2008,44(16):54-55,97.
ZHANG Congpin, LIU Chunhong, XU Jiucheng. Research on alpha-beta pruning of heuristic search in game-playing tree[J]. Computer Engineering and Applications, 2008, 44(16): 54-55, 97.

[2]刘知青,李文峰.现代计算机围棋基础[M].北京:北京邮电大学出版社,2011:63-80.

[3]GELLY S, WANG Yizao, MUNOS R, et al. Modification of UCT with patterns in Monte-Carlo Go[R/OL]. [2011-10-15]. http://219.142.86.87/paper/RR-6062.pdf.

[4]GELLY S, WANG Yizao. Exploration exploitation in Go: UCT for Monte-Carlo Go[C/OL]. [2011-10-15]. http://wenku.baidu.com/view/66c2edd6b9f3f90f76c61bc0.html.

[5]张汝波,周宁,顾国昌,等.基于强化学习的智能机器人避碰方法研究[J].机器人,1995,21(3):204-209.
ZHANG Rubo, ZHOU Ning, GU Guochang, et al. Reinforcement learning based obstacle avoidance learning for intelligent robot[J]. Robot, 1995, 21(3): 204-209.

[6]沈晶,顾国昌,刘海波.基于免疫聚类的自动分层强化学习方法研究[J].哈尔滨工程大学学报,2007,28(4):423-428.
SHEN Jing, GU Guochang, LIU Haibo. Hierarchical reinforcement learning with an automatically generated hierarchy based on immune clustering[J]. Journal of Harbin Engineering University, 2007, 28(4): 423-428.

[7]BAE J, CHHATBAR P, FRANCIS J T, et al. Reinforcement learning via kernel temporal difference[C]//Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Boston, USA, 2011: 5662-5665.

[8]SUTTON R S. Learning to predict by the methods of temporal difference[J]. Machine Learning, 1988, 3(1): 9-44.

[9]KAELBLING L P, LITTMAN M L, MOORE A W. Reinforcement learning: a survey[J]. Journal of Artificial Intelligence Research, 1996, 4: 237-285.

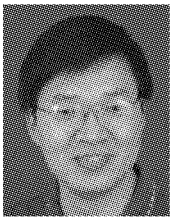
[10]阿培丁.机器学习导论[M].范明,昝红英,牛常勇,译.北京:机械工业出版社,2009:372-390.

[11]SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. Cambridge, USA: The MIT Press, 1997.

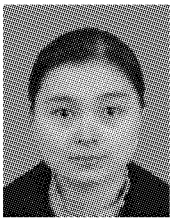
[12]聂卫平,冯大树.聂卫平围棋道场[M].北京:北京体育大学出版社,2004.

[13]徐长明,马宗民,徐心和,等.面向机器博弈的即时差分学习研究[J].计算机科学,2010,37(8):219-224.
XU Changming, MA Zongmin, XU Xinhe, et al. Study of temporal difference learning in computer games[J]. Computer Science, 2010, 37(8): 219-224.

作者简介:



张小川,男,1965年生,教授,中国人工智能学会机器博弈专业委员会副主任.主要研究方向为人工智能、人工生命、计算机软件等.主持国家级、省部级项目6项,横向项目30余项,曾获重庆市自然科学奖1项、科技进步奖1项,重庆市教学成果奖1项.主编教材3部,发表学术论文50余篇.



唐艳,女,1987年生,硕士研究生,主要研究方向为计算智能与智能软件.