Computer System Networking and Telecommunications
计算机系统网络和电信，2019,1(2)
ISSN:2661-3719 (Print); 2661-3727 (Online)

On the Basic Principle of Reinforcement Learning and Monte Carlo

Jiacheng XIE Zhengde BAO Yawen TANG
School of Computer and Software, Jincheng College, Sichuan University, Chengdu, 611731

Abstract

With the rapid development of the intelligent science, with the result of Alpha Go/ Zero, the prestige of the reinforcement learning RL is gradually enhanced, which is a kind of goal which can be dynamically selected and the optimal execution choice is obtained (the optimal solution is selected). So that the total bonus value of the final feedback reaches the maximum learning method. In an enhanced learning environment, there is a need for a dynamic, indefinite unit that can be tested in the entire environment mode and the correct execution selection is found in the context of using such a dynamic unit. Monterey In the Carlotree search algorithm, the multi-simulation of the problem and the prediction of the best next step based on the simulation results can be used to strengthen the learning algorithm. In this paper, based on the basic principle of the machine-intensive learning and the Monte-Carlo tree, the theory of combining the two with the field of artificial intelligence is briefly discussed.

Key Words

Machine Learning, Reinforcement Learning, Monte Carlo tree, Artificial Intelligence

DOI:10.18686/jsjxt.v1i2.700

浅析强化学习与蒙特卡洛树的基本原理

谢嘉诚 鲍正德 唐娅雯
四川大学锦城学院计算机与软件学院，四川成都，611731

摘要

如今智能科学快速发展，伴随着 Alpha Go/Zero 取得的成果，强化学习(Reinforcement Learning RL)的声望渐渐增强，这是一种能自主地进行动态选择，达到获取最优执行选择（选取最优解）的目的，使得最终回馈的奖励总值达到最大的学习方法。在强化学习的运行环境中，需求一种动态的不定单元，在使用这种动态单元的前提条件下，才能在整个环境模式中进行试验并发现正确的执行选择。蒙特卡洛树的搜索算法中，对问题的多次模拟以及基于模拟结果对最佳下一步的预测可用于强化学习算法。本文基于机器强化学习与蒙特卡洛树的基本原理，浅谈了关于将两者结合运用于人工智能领域的理论

关键词

机器学习；强化学习；蒙特卡洛树；人工智能

1.引言

在现代电子科技技术发展中，作为重点关注对象的机器强化学习带来的效益非常巨大。人工智能作为机器学习领域的研究方向之一，但仍在发展中受到了阻碍。如何在这个全新的领域取得有效进展是开发人员与研究者值得思考的问题。在能够体现人工智能技术的领域中，智能游戏有很高的代表性，它能够在简单的应用层上表现出一定的人工智能技术。研究技术的深入推动着更深层次领域的人工智能技术的开发，是一个由浅入深的过程。

2.机器强化学习

智能体在动态系统（环境）中以分析试验的方式进行学习，通过与动态环境进行信息交换获取优劣信号，

从而不断累积最优化执行选择（累积奖励值）的学习行为。[1]

2.1 强化学习的原理与基本模型

强化学习是从动物学习、参数扰动自适应控制等相关理论发展而来的。其基本原理是：若在某复杂的动态环境中学习系统的某一项执行决策，导致了环境向系统反馈回正向的奖励信号（优势强化信号），那么系统在继续运行时会优先执行此类决策以获得更多的奖励信号[2]。强化学习的目标是在每个不定的动态状态中发现最优选择，以使最终的有利奖励值（优势信号次数）总和最大。强化学习将学习行为认为是一种试探模拟得出结果并给与评价的过程，

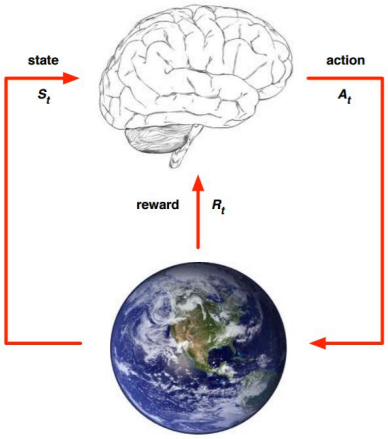
当学习系统在运行环境中任意执行一此操作以后，环境分析此次操作并作出相应的变化（环境变量在此时改变），同时反馈一个强化信号（优势或劣势信号）给学习系统，系统在识别反馈收到的强化信号（判断优劣）后，在决定下一次操作前会分析新的环境状态。而系统每确定一次执行选择，都会使环境回馈奖励强化信号的概率增大，使奖励总值不断增加。单次执行操作的结果所产生的影响不仅只决定此次的强化信号类别，也会改变下一次模拟试验时的环境状态和最终的奖励总值（即每一次模拟试验都会使得最新时刻的动态变量改变）。

强化学习系统执行试验的最终目的就是动态地分析并更改自身数据，以求取得奖励信号的最大化，即最优选择。

2.2 强化学习的简单模型设计

若用下图中的大脑代表强化学习中的算法执行载体，学习系统可以通过控制载体来实行决策，就是出优势选择（决定执行方向）。图中地球代表系统中的动态环境，环境拥有独立的随机动态模型，环境的状态（State）会随着我们随机的执行一次操作（Action）之后发生一定的变化，成为新的环境 State+1.同时我们会因为执行了一次操作 At 而得到环境回馈的延时奖励（Reward）Rt+1。在下一时刻运行时，系统可以通过原理分析继续选择下一个最优化操作 At+1,又得到一个新的环境状态 State+2,继而反馈得到新的延时奖励值 Reward+2。通过模型的不断运行我们会得到 n 个延时奖励值（Reward+1、Reward+2、.....Reward+n），而系统所执行的所有操作（n 次 At）都会使延时奖励值 Reward 的总和达到最大。

在这个简单模型中我们可以理解到关键的三个要素:执行操作(Action)、环境状态(State)和延时奖励(Reward)。在时间为 t 时整个系统的运行模式是这样的:S(t)+A(t)=R(t+1)。说明了最基本的强化学习运行原则。



3.蒙特卡洛树搜索(随机抽样或统计试验方法)

一种利用随机数生成用于模拟变量取值规律的数学原理方法。在计算机领域作为一种算法思路:蒙特卡洛搜索算法(MCTS)在机器强化学习领域中有着重要运用。

3.1 基本原理思想

蒙特卡洛方法的基本理念是在面向求取含有未知随机变量的概率事件时，通过某种“试验”的方法以得出所包含的随机变量的均值或某种情况（特殊事件）成立的概率，作为此类问题的近似解。这种方法通过分析事件概率或随机变量的变化特征，利用数学方法确定一个基准模型（即变化规律）用于模拟，按照模型规律对问题变量进行捕捉并求解。通常利用蒙特卡洛方法解决问题的顺序如下：第一步：构造或描述概率过程，即正确构建概率模型并确定运算方法；第二步：需从已确立的概率模型分布中进行随机抽样；第三步：建立估计量即确定一个随机变量作为问题的解[3]。

而在 MCTS 中,蒙特卡洛树搜索在制定最优解的执行方案前，会预先进行多次试验性博弈，并根据每次试验得出的结果不断分析以更新博弈树中的数据以及调整自身参数。蒙特卡洛树搜索的主要理念是搜索，其含义是博弈树中由根节点作为起点，到终节点结束的一组试验集合，路径是由当前环境状态（根节点）到任意一个未被选择过的节点，直到最后一个节点（终节点）。而在遇到未完全访问节点时，系统则会选取未被选择的

子节点进行试验以保证每一次的试验路径上至少有一个未被选取过的节点，以避免重复试验。在得出一次模拟结果后，信息将被反馈至当前环境状态下的根节点，且路径上的所有节点将会分析数据并更新自身信息（用于判断下一次选择），当根节点以下的子节点全部试验结束后，系统则会根据收集的信息（优劣信号次数）决定下一步的执行选择。

MCTS 的基本原理可分为四点:

1)选择：从当前环境状态量（选为根节点）开始，按照预先设定的系统选取规则，提取余下所有子节点。

2)扩展：由当前子节点扩展一个或多个符合系统约定的下一级子节点。

3）模拟：面向待选取的子节点采取随机的模式进行一定次数的模拟试验，直到在终节点完成模拟后，由根节点得出此组模拟所得的奖惩值（优劣对比）。

4)结果回传：在某一子节点经过多次模拟试验得出奖惩值后，覆盖更新此节点的试验次数与奖惩值。并将数据回传至其所有更高一级节点并更新路径上的所有节点的数据信息。

3.2 投入领域

蒙特卡洛树依靠其原理及运算方法投入了众多面向复杂动态问题的领域。运算次数即“试验”次数越多,所求得的解则越接近最优解的思想在这些面向大量数据样本的领域中发挥着重要的作用。

在对于金融工程学的核心目标而言，蒙特卡洛树就能能得以很好的运用，金融工程学其目的是为了在风险不确定情况下利用创新金融工具，以求更有效地分配和再分配个体所面临的各种经济风险，以优化它们的风险/收益率。而在金融领域中金融银行、证券投资、商业公司等载体面对已知情况下的选取下一步发展趋势具有不确定性与风险性，需要通过大量的试验分析市场情况以求取更稳定的执行选择。而蒙特卡洛树则可与之结合以加快问题分析和避免掉冗杂的数学报道和演算过程，能够在复杂的动态环境中筛选出最大利益化的执行选择。并且此类运用不仅仅只局限于金融与财务领域，在企业管理、商品定价、专利权价值估算等方面也取得了不小成就，使得原本相当复杂问题的解法变得简单快速。

4.强化学习与蒙特卡洛树的运用

4.1 运用方向

强化学习与蒙特卡洛树的集合运用主要体现在人工智能领域，而在人工智能领域中智能医疗对于深度强化学习实际应用具有很高效益。通过强化学习的方法可以对病患不同时间节点的身体状况和病理反应作出分析,类似于 Alpha Go 的运行模式,在少量的初始样本上得出大量的模拟试验数据,以得出最优解决方案。[4]

但在游戏领域中，虽然通过深度强化学习技术与 Alpha Zero 的成功实例解决了启发式搜索问题等难题，却因并不具有泛化性能而导致不能在各方面普遍的适用。而更核心实时性需求、态势感知与估计、非完全信息博弈等复杂性系统问题利用当前层次的深度强化学习技术尚未能有很好的突破。

4.2 实例分析(Alpha Go/Zero)

4.2.1 应用模块

目前,在机器强化学习与蒙特卡洛树的运用中,Alpha Go/Zero 是具有很强代表性意义的实例,其核心部分就包括了蒙特卡洛树搜索(使用 PUCT 函数的一种树遍历的特定变体)与强化学习(通过自我对局来训练网络)。

4.2.2 结合 MCTS 的执行方式

其中蒙特卡洛树搜索(MCTS)的应用主要是用于帮助 Alpha 在所有的执行选择中捕捉到当下的最优解。依靠的是其中的模拟、反向传播、数据记录更新等方法进行分析。在 Alpha 所进行的“试验”中,将每一步的运行路线及其结果描绘出来,就能得到一副树状图(博弈树或游戏树)。以井字棋为例,当对手下了第一步棋之后,在当前状态开始进行一种动态模拟(即当前有八种选择方式),模拟的作用即任意挑选一种路径(节点)并执行,直到最终结果(终端节点),在简单应用中模拟只是一串从当前状态到终端结果的移动序列。在模拟执行完当前序列的所有“试验”后得出一个执行结果(优或劣),系统将所得执行结果反馈回模拟开始节点(根节点)实现反向传播,并在各个节点记录模拟次数和优劣情况。当此根节点下端所有节点路径被模拟完成之后,在选取下一节点时,系统会根据下一等级节点的模拟次数与优劣情况作出选择以确定下一时间的根节点。例如在两个节点 A 与 B 中,A 与 B“试验”同等次数,A 节点所得优势信号多于 B

节点,则会确定下一节点为 A 节点[5]。

4.2.3 利用强化学习完善网格

而在具体的对弈过程中,对手的下一步是不可预判的,所以 Alpha 采用了强化学习的方法,通过自我对局来进行模拟“试验”。而在对弈中 Alpha 不依靠实时环境进行分析,而是通过强化学习将所有的可能“试验”结果进行模拟,依据对手的下一步作出选择节点的更改。

4.2.4 计算能力与试验模式的局限

但基于强化学习与蒙特卡洛树的原理以及基本形式,使得其在不够大的博弈环境中给与不够充足的运算时间也无法得出确定的最优执行选择,在根节点的分析时对于最优解的执行选择并不能依据足够多试验结论来给出合理的判断。而在一些复杂的研发领域中,因为 MCTS 搜索需要达到一定数量上的迭代以后才能得出一个靠近一个合理的近似解，而往往系统的计算能力与可维持的行动时间不能满足于进行面对大量数据的模拟试验。这两个方面的局限造成了模拟实验少量不精确、多量超负荷的瓶颈，即并不适用于小型博弈环境与少量运算问题解，而在求取复杂问题时,计算能力的强弱与系统可维持的行动时间是关键因素。

5.结束语

在如今的科研领域,机器学习并不是最受关注的一个方向,但却是最能改变现代生活、科技以及发展的一个领域。强化学习作为一种重要的机器学习方式对于未来智能科学的发展不可或缺,但因为机器学习的可解释性与可干预性并不强,所以也对此类研究造成了一定的发展阻碍。在未来的发展中,让机器学习行为具有可解释性是关键所在,如何避免潜在危机而研发有利于人类发展的科学技术是当前最需要解决的问题。

参考文献

[1]王鹏程. 基于深度强化学习的非完备信息机器博弈研究[D].哈尔滨工业大学,2017.
[2]黄炳强,曹广益,王占全.强化学习原理、算法及应用[J].河北工业大学学报,2006(06):34-38.
[3]李承奥.基于机器强化学习与蒙特卡洛树的基本原理及其应用[J].通讯世界,2019,26(02):212-213.
[4]许杰. 基于机器学习的医疗健康分类方法研究[D].郑州大学,2018.
[5]林云川. 基于深度学习和蒙特卡洛树搜索的围棋博弈研究[D].哈尔滨工业大学,2018.

作者简介

第一作者：谢嘉诚（1999-），男，汉，四川省乐山市，专科，四川大学锦城学院。
第二作者（通讯作者）：鲍正德（1989-），男，汉，黑龙江哈尔滨，研究生，四川大学锦城学院，研究方向：电子商务。
第三作者：唐娅雯（1999-），女，汉，四川省资阳市，本科，四川大学锦城学院，研究方向：信息管理、J2EE