基于机器强化学习与蒙特卡洛树的基本原理及其应用

李承奥(中国科学技术大学少年班学院,安徽 合肥 230026)

【摘 要】强化学习(Reinforcement Learning RL),又称增强学习,是一种重要的机器学习方法,它不像监督学习需要先验知识,而是通过不断与环境交互来获得知识,自主地进行动作选择。本文通过对强化学习和蒙特卡洛方法的基本原理的介绍,系统地探讨了其在金融投资、机器人控制、医疗等方面的应用。对强化学习的应用推广有创新性启发。

【关键词】强化学习;蒙特卡洛树;金融投资;医疗
【中图分类号】TP18 【文献标识码】A 【文章编号】1006-4222(2019)02-0212-02

1 背 景

继 Open AI 在 5v5 DOTA 2 中战胜人类玩家后不久,DeepMind 又在多智能体学习方面取得新的进展,其最新的工作展示了智能体在复杂的第一人称多人游戏《雷神之锤 3 竞技场》中达到人类水平,并能与人类玩家合作。这一新的突破证明了强化学习在人工智能体学习竞技游戏中的巨大成功。然而,利用强化学习突破竞技游戏并不是我们的最终目的,将这一新的机器学习方式应用到社会生活中并创造更大的价值才是其应当具有的作用。

2 强化学习

2.1 强化学习的基本原理

强化学习是智能系统从环境到行为映射的学习,它是一种试探评价过程。Agent 选择一个动作作用于环境,环境接受该动作后状态发生变化,同时产生一个强化信号反馈给 Agent,Agent 根据强化信号和环境当前状态再选择下一个动作,选择的原则是使受到正强化的概率增大。选择的动作不仅影响立即强化值,而且影响环境下一时刻的状态及最终强化值。

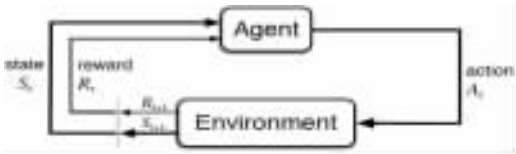

图 1

强化学习不同于传统机器学习中的监督学习,主要表现在强化信号上,强化学习中由环境提供的强化信号是对 Agent 所产生动作的好坏的一种评价,而不是告诉 Agent 如何产生正确的动作。由于外部环境提供了很少的信息,Agent 必须靠自身的经历进行学习。通过这种方式,Agent 在行动——评价的环境中获得知识,改进行动方案以适应环境。

2.2 强化学习的应用

2.2.1 强化学习在金融投资中的应用

在传统的监督学习中,机器需要从有标记的训练数据中推导出预测函数。以对银行股收盘价格的预测为例。监督学习中采用银行股在第 n 天的开盘价、收盘价、最高(低)价、交易量等作为输入数据,第 n+1 天的收盘价作为输出数据,进行训练,从而预测收盘价格;而在强化学习中,使用银行股在第 n 天的开盘价、收盘价、最高(低)价、交易量等数据作为输入,而将交易员在第 n 天的现金、持仓价值、持仓量等数据作为 State,使用交易指令:买入、卖出、持有等作为 Agent 的 Action,来预测收盘价格。

目前,强化学习模型在金融领域的实验还很少,所以并无明确数据表明强化学习在金融领域的预测效果要优于监督学习。但是,由于强化学习综合利用了交易当天与环境有关的数据,所以具有超过监督学习的潜力。

2.2.2 强化学习在机器人控制中的应用

利用强化学习来控制机器人执行指定动作时,优势在于可以在缺少地面实况模型的情况下学习并完成任务,操作者只需设计合适的奖励函数,智能体就会在随意操作中触发正确的操作,从而获得奖励,不断学习,最终产生正确的操作,获得需要学习的能力。但问题在于强化学习往往需要很长的时间,即使可以通过优化奖励函数来缩短学习时间,但由于设计良好的奖励函数本已不易,提高学习效率就更加困难。如果地面实况模型已知,那么用强化学习来引导机器人行走就远不如针对模型进行规划效果好。所以,在常见的家用机器人领域,强化学习并非理想的学习方式,而在缺乏地面模型的情况下,如对金字塔或陵墓的发掘,对敌人建筑的探索时,或是在没有任何先前经验时,运用强化学习可能会有较好的效果。然而在这些情况下多是人为操控机器人进行操作,而在未知地面模型的情况下,如何设计奖励函数都是非常困难的问题。由此可见,虽然强化学习在机器人控制领域有一定的应用前景,但如何设置奖励函数从而提高学习效率实在是非常棘手的问题。

虽然从理论上讲,一个强大的强化学习系统能解决任何问题,但是,很容易看出,强化学习是存在其弊端。主要在于其奖励函数难以设计、效率不易提高,并且在众多情况下,强化学习的效果不尽如人意,甚至不如监督学习。因此,如何改进强化学习的方式来优化其效果值得思考。

3 蒙特卡洛树搜索

3.1 蒙特卡洛树的基本原理

蒙特卡罗树搜索是一种行动规划形式。它包含四步:选择,拓展,模拟,反向传播。在开始阶段,搜索树只有一个节点。搜索树中的每一个节点包含了三个基本信息:代表的局面,被访问的次数,累计评分。

3.1.1 选择(Selection)

在选择阶段,需要从决策局面 R 出发向下选择一个最需被拓展的节点 N,在局面 R 下,对于所有可行动作都已经被拓展过的节点,我们使用 UCB 公式:

$$UCB1(S_i)=\bar{V}_i+c\sqrt{\frac{\log N}{n_i}},c=2$$

计算其所有子节点的 UCB 值,并选择值最大的子节点反复迭代;对于有未被拓展过的可行动作的节点,该点即是目标

节点 N;对于已结束的节点,直接进行 Backpropagation。每个被检查节点的被访问次数在该阶段都会增加。在反复迭代后,在底端将找到一个节点,来继续之后步骤。

3.1.2 **拓展**(Expansion)

对于最需被拓展的节点 N 及其未被拓展的动作 A,创建新的节点 Nn 作为 N 的一个新子节点。Nn 的局面就是节点 N 在执行了动作 A 之后的局面。

3.1.3 **模拟**(Simulation)

从 Nn 开始运行一个模拟的输出,直到博弈游戏结束。

3.1.4 **反向传播**(Backpropagation)

在 Nn 的模拟结束之后,从根节点到 N 的路径上的所有节点都会根据本次模拟的结果来增加自己的累计评分。

随着迭代次数的增加,搜索树的规模也不断扩大。在达到一定时间或迭代次数后结束,选择根节点下的最优子节点作为决策结果。

3.2 蒙特卡洛树的特点

蒙特卡洛树搜索是一种广泛运用于强化学习的决策方法,它的本质在于从当前的根结点下的子节点中选择最优子节点来作为决策的结果,即采用一种贪心算法使每一步的累积计分都达到最大,从而使得到最好的结果。但是,蒙特卡洛算法只能解决有限步数的问题,当问题步数太多或无限时,由于迭代次数不允许,无法用蒙特卡洛算法来解决。例如在解决围棋问题时,由于围棋的状态空间太过庞大,利用暴力搜索的方法无法穷尽。所以只能使用经过改进的 Monte-Carlo 法,即从给定落子位置开始,随机采样,得到一个模拟的结果。经过多次采样之后,将平均成功率返回作为该点的成功率。

4 AlphaZero 的自主学习及其启发

AlphaGo 是第一个战胜围棋世界冠军的人工智能机器人,由 Google 旗下 DeepMind 公司戴密斯·哈萨比斯领衔的团队开发。2016 年 3 月,AlphaGo 在韩国首尔挑战世界围棋冠军李世石,最终 AlphaGo 以 4 比 1 的总比分取得胜利。2017 年 5 月,在中国乌镇围棋峰会上,AlphaGo 的强化版本 AlphaGo-Master 以 3 比 0 的总比分战胜排名世界第一柯洁。

2017 年 10 月,最强版阿尔法围棋 AlphaZero 经过短短 3d 的自我训练打败了旧版 AlphaGo。经过 40d 自我训练,AlphaZero 又打败了 AlphaGoMaster 版本。

DeepMind 团队在创造与改进 AlphaGo 的过程中,采用了深度卷积网络、Monte-Carlo、Rolloutpolicynetwork、强化学习策略网络、价值网络等算法。而在实际对弈过程中,只用了 Rolloutpolicynetwork(简单人工提取的棋局特征+线性 softmax 分类器)和价值网络(是不同于策略网络的一个评价体系,采用回归的方法,估计当前状态的获胜期望,是对于当前状态所有动作情况直到结束的一个考虑)。

AlphaZero 则完全独立地通过自我博弈强化学习来完成训练。①随机博弈开始就没有任何监督或人工数据。②它只使用棋盘上的黑白子作为输入特征。③只使用一个神经网络,而不是分开的策略网络和价值网络。④只使用依赖于单一神经网络的简化版树搜索来评估落子概率和落子对局势的影响,不再使用蒙特卡洛方法。

AlphaGo 对强化学习与蒙特卡洛方法的应用以及 AlphaZero 的自主学习都给我们以蒙特卡洛方法应用的极大启发。

4.1 强化学习与蒙特卡洛方法在医疗中的应用

AlphaGo 中所蕴含的机器学习的技术目前即将用于智能医疗,据报道,为实现“将 AlphaGo 和医疗、机器人等结合”的计划,哈萨比斯于 2016 年初在英国的“巴比伦”公司投资了 2500 万美元。巴比伦正在开发自动搜索医疗信息、寻找诊断、推荐最佳后续化验和检查项目的人工智能 APP。如果 AlphaGo 和“巴比伦”结合,诊断的准确度将得到划时代性提高。

在蛋白质工程技术中 AlphaGo 运用的蒙特卡洛方法也可发挥相应作用,由于蛋白质工程需先设计决定蛋白质性质的基因,若采用蒙特卡洛树的方法,可将决定一个氨基酸的三个碱基构成的密码子看成一个节点,在已知某些天然人工蛋白质的性质时,将此作为已知数据来缩小蒙特卡洛树的规模,从而尝试构造出具有全新性质的蛋白质,来进行基因工程实验。

4.2 强化学习与蒙特卡洛方法在组合优化中的应用

在工业生产中经常有与组合优化有关的流程,例如生产工序的安排,人员调度,财物分配,等。蒙特卡洛方法在游戏中的应用实际上是于博弈中的应用,而组合优化在一定程度上也可看作是单人的博弈。每个过程可看作是一个节点,节点的性质包括进行该过程所需时间与该过程开始时间,利用蒙特卡洛方法结合强化学习就可找到奖励函数值最大也即效率最高的方案。据悉,AlphaGo 团队将与英国电力能源部门合作来提高能源利用效率,这其实就是蒙特卡洛方法在组合优化中的应用。

曾有人提出蒙特卡洛方法用于星际飞行。这本质也是在组合优化中的应用。若将星际飞行看作是访问多个行星的旅程,可以利用这些行星的引力场来节省燃料。问题在于确定到达和离开每个行星表面或轨道的时间。可将这个问题看作一个决策树。如果将时间分成区间,在每个行星处都要进行一次决策:在何时到达,又在何时离开。每一选择都会影响之后的选择。首先,不能在到达前离开;其次,之前的选择决定着当前剩余燃料以及所处的位置。于是这一问题就可以运用蒙特卡洛方法结合强化学习来解决。

5 总结

强化学习是机器学习中一种新奇的方法,而蒙特卡洛树是广泛应用于强化学习的一种决策方法。这些方法注定会在日后的应用中大放异彩。本文简要介绍了两者的基本原理并给出了应用思路,虽然距离实际应用仍有很长的距离,但这些创新性的思维依然对强化学习与蒙特卡洛树的应用有启发意义。

参考文献

[1]张汝波,顾国昌,刘照德,王醒策.强化学习理论、算法及应用[J].控制理论与应用,2000,10,17(5):637~642.

[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.

[3]周志华.机器学习[M].北京:清华大学出版社,2016,1.

[4]邱 强.要不我也说说 AlphaGo? [EB/OL].2017-5-27.

收稿日期:2019-1-20