# 中国人民大学附属中学

# 2019-2020 学年度高二研究性学习课题文献综述

课题编号：

课题名称：

课题成员：

指导教师：

# 1 强化学习简介

## 1.1 定义

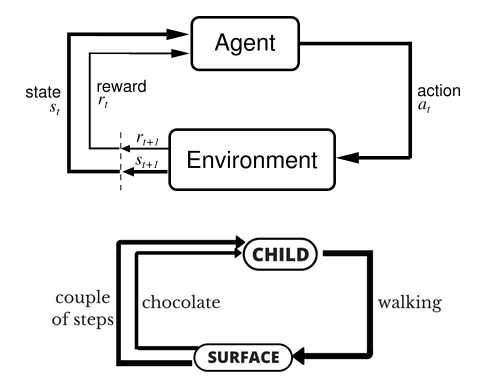
强化学习是机器学习的一个重要分支，是多学科多领域交叉的一个产物，它的本质是解决 **decision making 问题**，即自动进行决策，并且可以做连续决策。

它主要包含四个元素，**agent**，环境状态，行动，奖励, 强化学习的目标就是获得最多的累计奖励。

让我们以小孩学习走路来做个形象的例子：

小孩想要走路，但在这之前，他需要先站起来，站起来之后还要保持平衡，接下来还要先迈出一条腿，是左腿还是右腿，迈出一步后还要迈出下一步。

小孩就是 agent



* 强化学习是用来做决定的工具
* 要控制的对象是agent
* 强化学习的目标是获得更多奖励

## 1.2 分类

### 1.2.1不理解环境（Model-Free RL）和理解环境（Model-Based RL）

不理解环境（Model-Free RL）和理解环境（Model-Based RL）。其区别为是否对环境进行建模。理解环境拥有想象力，可以在模拟的环境中进行试验，预判出最优策略，再把这一策略应用到真实环境中。

### 1.2.2 基于概率（Policy-Based RL）和基于价值（Value-Based RL）

基于概率在感官出入信号后，根据每种动作的概率来选择将要做出的动作。而基于价值直接选择价值最高的动作。基于概率的方法可以在一个连续的动作中根据概率分布来选择一个特定的动作，而基于价值的方法不行。

基于概率和基于价值还被结合起来，成为了Actor-Critic算法，它根据概率选择动作，再为动作评价价值。

### 1.2.3 回合更新（Monte-Carlo update）和单步更新（Temporal-Difference update）

回合更新在游戏开始后进行一系列操作，在游戏结束后才更新游戏策略。单步更新在每一个动作执行完后就直接更新策略。单步更新更有效率。

### 1.2.4 在线学习（On-Policy）和离线学习（Off-Policy）

在线学习必须边玩边学，离线学习可以边玩边学，也可以看着别人玩来学习。

# 2 算法

## 2.1 Q Learning

Q Learning需要一张Q表，用来记录每一个状态Si和行为Ai对应的奖励值，记为Q(s, a)。每一种状态时要进行什么动作只需要查Q表就能知道了。那么如何更行Q表呢？

假设现在已知Q(s1, a1) Q(s1, a2)。在s1状态后进行了a1的操作，变成了s2状态。然后在s2状态下尝试a1,a2两种动作，得到奖励值，于是就有了Q(s2, a1), Q(s2, a2)。下面用Q(s2)来更行Q(s1, a1)。设到达s2时获得的奖励值为r，s2状态下进行的所有a得到的奖励值中最大的为maxQ(s2)。把Q(s1, a1)更新为Q(s1, a1) + α \* ( r + y \* maxQ(s2) – Q(s1, a1) )。现实意义为把原价值加上新价值和原价值的差值，α是学习效率，y是衰减值（未来的奖励有多重要）。

## 2.2 Sarsa

Sarsa和Q Learning非常相似，都使用了Q表，只有状态转移不一样。Sarsa并不尝试动作，而是直接采取行动，所以Sarsa的状态转移方程为Q(s1, a1) = Q(s1, a1) + α \* ( r + y \* Q(s2, a1) – Q(s1, a1) )。

Q learning倾向于选择最有可能成功的路线，而Sarsa倾向于选择危险最少的路线。Q learning是离线学习，Sarsa是在线学习。

## 2.3 Sarsa(λ)

在Sarsa的基础上，Sarsa(λ)记录每回合走过的路线，λ是脚步衰减值，离奖励越近的动作受到更新的力度越大。

Sarsa(λ)是回合更新，Sarsa是单步更新。

## 2.4 DQN

传统的Q表法不能适应现在的机器学习任务，因为围棋等游戏的状态数太多，Q表存不下来。于是DeepMind首先使用了神经网络代替了Q表。

状态转移方程为：新NN = 老NN + α(Q现实 – Q估计)

DQN还有两种优化方法：Experience replay and Fixed Q-targets。它们分别是学习原来的经历和重复利用很久以前的参数。

# 3 应用实例

## 3.1 OpenAI Dota2 Rerun

Dota2是一个电子游戏，在人工智能OpenAi Five面前，人类的Dota2冠军队Team OG不堪一击，现在，OpenAI的新人工智能Rerun全面碾压OpenAI Five，胜率达到98%。Rerun中就应用了强化学习技术。

Dota2比围棋更复杂，围棋一局大约要走150步，而Dota2一局比赛要dui 电脑做出超过20000次操作。而且由于战争迷雾的存在，Dota2中，只能看到局部的信息，其他的信息都处于隐藏的状态。

训练使用的是扩展版的近端策略优化（PPO）方法，这也是OpenAI现在默认的强化学习训练方法。这些智能体的目标是最大化未来奖励的指数衰减和。在训练策略的过程中，OpenAI Five没有用到人类游戏数据，而是通过自我博弈。在围棋、象棋等问题上，也应用了类似的方式训练。其中，80%的战斗中对手是使用了最新参数的分身，而20%的对手是老参数的分身。每经过10次迭代之后，新训练出的分身就被标为老前辈。如果当前正训练的AI击败了新秀或者老前辈，系统就会根据学习率更新参数。按照OpenAI CTO此前的说法，击败OG前OpenAI Five已经练习了相当于45000年Dota。AI每天的训练量相当于人类打180年游戏。

从这个应用例子中可以看出，用强化学习玩对战类的游戏，可以让电脑自我对弈，其中对手使用上一次训练的分身。此外，还可以人工写一些脚本和电脑对战，让电脑学习人类游戏的经验。

## 3.2雷神之锤III竞技场

继OpenAI之后，DeepMind也在多智能体强化学习方面秀肌肉：首次在第一人称射击游戏的多人模式中完胜人类，而且没有使用大量训练局数，轻松超过了人类水平。

从实验结果看，他们的AI已经超越了人类：在与由40个人类玩家组成的队伍对战时，纯AI的队伍完胜纯人类的队伍（平均多抢到16面旗），并且有95%的几率战胜AI与人混合组成的队伍。这个AI名叫“为了赢”（For the Win，FTW），只玩了将近45万场游戏，理解了如何有效地与人和其他的机器合作与竞争。

智能体必须从头开始学习如何在不可见（unseen）的环境中观察、行动、合作和竞争，所有这些都来自每场比赛的一个强化信号：他们的团队是否获胜。这是一个具有挑战性的学习问题，它的解决方法基于强化学习的三个一般思路:

* 我们不是训练一个智能体，而是训练一群智能体，它们通过组队玩游戏来学习，提供了多样化的队友和敌手。
* 群体中的每个智能体都学习自己的内部奖励信号，这些信号使得智能体能够产生自己的内部目标，例如夺取一面旗子。双重优化过程可以直接为了获胜优化智能体的内部奖励，并使用内部奖励的强化学习来学习智能体的策略。
* 智能体在两个时间尺度上运行，快速和慢速，这提高了它们使用内存和产生一致动作序列的能力。

从以上材料看出，强化学习的功能十分强大，用其训练出的智能体甚至可以有团队合作的能力。这也提示我们，训练时可以使用多个智能体一起训练，效果可能比只训练一个庞大的智能体效果更好。

# 5 结束语

强化学习是一 种无导师的在线学习技术。在马尔可夫环境中，最优搜索型强化学习算法已经被证明收敛性。但对非马尔可夫环境可以进一 步分为部分感知强化学习、函数估计、多agent强化学习以及强化学习偏差技术研究。目前对非马尔可夫环境下的强化学习研究正成为研究的热点。再加上现在热门的神经网络，可以把强化学习和神经网络结合达到更高级的能力。在棋类、游戏游戏等方面有很多应用。