强化学习、神经网络与多智能体中博弈论的应用

康子熙 2200017416

2023 年 6 月 8 日

摘要

当下，多智能体与博弈论的结合是一个非常热门的话题。在本学期的学习中，我选修了李文新老师的《人工智能基础》课程、朱老师为通班开设的《人工智能初级研讨班》课程、以及社会学系陶林老师的《社会博弈论》课程。我在这三门课程上分别吸纳了一些人工智能、多智能体以及博弈论的基础知识。在学习这三门课程的过程中，我惊讶的发现三门课程有很多相通的知识点以及逻辑。在这篇文章中，我会尝试着使用我在这三门课程上学到的知识，使用理论知识探讨博弈论在强化学习、神经网络和多智能体中的应用，并考虑其对于通用人工智能（AGI）实现的可能影响。

首先，我将从强化学习的角度出发，通过分析未来收益的估计，阐述如何利用重复博弈的模型来解释和优化强化学习的收益函数。然后，我会转向神经网络，特别是对生成对抗网络（GAN）的理解。在这一部分，我将可视化 GAN 中的分类器和生成器如何在无限重复博弈的环境中互相博弈进化，以此揭示 GAN 能获得优良生成效果的原因。

在多智能体的讨论中，我会关注博弈论如何帮助解决多智能体系统中的竞争和合作问题，同时也指出博弈论在多智能体系统中的限制和挑战。在总结部分，我们会对现有研究的局限性及未来研究方向进行概述，并提出 AGI 是否可以通过多智能体训练以及博弈论的辅助来实现的问题，以及 AI 的社会性如何通过博弈论得到实现。

关键词：多智能体；生成对抗网络；多任务学习；强化学习；重复博弈；贝叶斯博弈；

1 前言：背景介绍

人工智能（AI）的发展历程源远流长，从 20 世纪 40 年代的逻辑理论机，到近年的深度学习和神经网络，AI 已经取得了显著的进步。然而，AI 的发展并非独立的，它与博弈论，多智能体，强化学习等领域紧密相连。

博弈论最早由约翰·冯·诺伊曼和奥斯卡·摩根斯特恩在 1944 年提出，作为研究决策过程中相互影响的因素的数学理论。它的出现对 AI 的发展产生了深远影响，尤其在强化学习和多智能体的研究中。

多智能体系统（MAS）是自 1970 年代以来的一个研究热点，这种系统中包含多个能够互相沟通、协作或竞争的智能体。在 MAS 中，多个智能体需要协作或竞争以实现各自或共同的目标。强化学习和博弈论的结合为处理 MAS 中的问题提供了新的思路，也为 AI 的发展开辟了新的道路。

强化学习则是自 1980 年代起开始发展起来的一个领域，它涉及到智能体如何通过与环境的交互来学习和优化自己的行为。在强化学习中，博弈论的概念被用来解释智能体如何在不确定的环境中作出最优的决策。

本文将基于《人工智能初级研讨班》课程上邓小铁老师对多智能体的讲解、杨耀东老师对强化学习的讲解，从强化学习、神经网络和多智能体三个方面，探讨博弈论在 AI 中的应用。

2 强化学习中博弈论的应用

2.1 强化学习的训练：寻找子博弈完善均衡

在《人工智能基础》课程上，对一个强化学习的问题，我们可以这样建立一个问题的模型：

1. 初始状态:S0
2. 当前玩家:C
3. 动作:A
4. 状态转移函数:P
5. 终止状态:ST
6. 奖励函数:R

强化学习的目标则是在每一个状态 S，通过训练，或是用值函数表，或是用神经网络，通过这类方法获得一个针对于 S 的最优动作 a。在每一个 S 得到 a 并且继续重复行动的过程，就是一个经典的重复博弈的模型。

在知道以上强化学习的定义后，我们可以如下**抽象表述**一个简单的强化学习的模型：

1. 玩家：博弈者 a、博弈者 b（我们不妨设模型只有两个博弈者，多个博弈者的问题模型同理即可）
2. 策略：博弈者 a 在每一个属于自己的信息集 IA_i 上都拥有一个动作的集合 a_i；博弈者 b 在每一个属于自己的信息集 IB_j 上都拥有一个动作的集合 b_j。
3. 收益：在博弈树的叶节点，可以获得一个二维的收益值，分别表示博弈者 a 和博弈者 b 的收益。即收益函数 $u(va, vb)$。

强化学习的目标在这个模型中则变成了使 a、b 均可以变成更理性的决策者，选择对自己最有利的策略。

我们再观察一下值迭代和策略估值的过程，可以发现，在这两个过程中，迭代/估值的截止条件都是再进行一次迭代之后新策略和旧策略的差值 delta 小于预先设定的一个超参数 。**使用博弈论的角度判断，它们都是在寻找一个子博弈完善均衡（subgame perfect equilibrium）。**

这个结论是很容易证明的，强化学习的理想目标正是双方在面对全部状态时都没有**偏离最优策略的倾向**。而子博弈完善均衡在动态博弈中正是满足这一条件的策略组合。所以，对于强化学习问题，在不考虑历史的影响下，如果我们可以构建一个策略组合 $(a_1a_2..., b_1b_2...)$ 成为博弈的一个子博弈完善均衡，那么对于每一个信息集我们都可以构建出一个最优反应映射，从而使得我们的强化学习问题得到解决。

计算机科学家们是如何利用这样的结论的呢？**对于小规模问题，使用逆推归纳法，一定可以获得子博弈完善均衡** [1] ，即使用蒙特卡洛 + 值函数表的方法存储子博弈完善均衡对应的反应映射；对于大规模问题，我们没有空间去存储子博弈完善均衡对应的反应映射，但是我们可以使用神经网络来近似这个反应映射，从而使得我们的强化学习问题得到解决。

通过这个例子，我们可以看出来，智慧的计算机科学家们这样设计强化学习的训练方法是非常科学的，我们也可以通过博弈论的证明来说明这种训练方式得出的结果是最优的。在知道强化学习方法的合理性之后，科学家们如何能够加快训练速度、提高神经网络的预测精度是现在最热门的方向之一。随着我们对动态博弈的理解加深，显然，博弈论会为强化学习的训练提供更多的帮助。

2.2 强化学习的训练：寻找合适的收益

在分析了强化学习方法的合理性之后，我们希望在训练的过程中也应用到博弈论的方法辅助进行训练。其中一个可以使用博弈论优化的部分便是强化学习中 AI 所获得的收益。收益与 AI 的策略是直接相关的，我们可以通过控制收益相关参数控制 AI 的策略。

那么，在强化学习的蒙特卡洛方法中、值迭代中、策略估值中，我们应该如何控制收益函数来优化 AI 的策略呢？如果我们对训练的 AI 有特殊的需求，我们又该怎么控制收益函数来控制 AI 呢？

2.2.1 累积收益

在《人工智能基础》课程上，我们学习了强化学习的收益函数，即累积收益。在强化学习的领域里，我们这样定义累积收益 G[2]：

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{1}$$

这个式子为什么是合理的？我们可以从博弈论的角度来分析这个问题。

用《人工智能基础》课程第四次作业井字棋举例。我实现了一个使用蒙特卡洛和值迭代的井字棋 AI。在使用强化学习方法训练时，我们假设井字棋的搜索量是很大的，我们必须使用估计每个状态的 Value 的方法来寻找子博弈完善均衡。于是我们可以通过回归更新每个节点的 Value。而我们使用的更新函数正是 G_t。

使用有限重复博弈的思想分析井字棋问题。我们可以将任意一方放置一枚棋子的过程视为有限重复博弈的一个阶段。那么累积收益就是有限次重复博弈中博弈者 a 所得到的收益流。当博弈进行了 n 个阶段时，我们定义该博弈的平均收益为 $v = \frac{G}{n}$，也就是说，**强化学习的蒙特卡洛方法正是使用收益流寻找最大平均收益的过程。这便说明了强化学习蒙特卡洛方法的合理性。**

2.2.2 冷酷触发：更好的 AI

我们现在考虑这样一个扩充军备的游戏：

1. 有两个玩家，分别是博弈者 a 和博弈者 b。

2. 博弈者 a 有两个动作：扩军 c 和不扩军 d。

3. 博弈者 b 有两个动作：扩军 c 和不扩军 d。

4. 博弈者 a 和博弈者 b 的收益矩阵如下：

	c	d
c	-5,-5	5,-10
d	-10,5	1,1

我们可以看到，这个游戏的收益矩阵是一个非常经典的囚徒困境的收益矩阵。在这个游戏中，博弈者 a 和博弈者 b 都有一个最优策略，那就是扩军。但是，如果我们让博弈者 a 和博弈者 b 都使用最优策略，那么他们的收益就会是 (-5,-5)。这个收益显然不是最优的，因为如果他们都选择合作，那么他们的收益就会是 (1,1)。

如果我们使用强化学习来训练博弈者 a 和博弈者 b，那么他们的收益就会是 (-5,-5)。这是因为，如果我们使用累积收益作为收益函数，那么博弈者 a 和博弈者 b 都会选择扩军，因为这样他们的收益会更高。但是在现实生活中，我们希望我们的 AI 具有**道德性**，或者**合作性**。我们希望我们的 AI 能够选择合作，而不是内卷。

那么，我们应该如何训练我们的 AI 呢？我们可以使用博弈论中的**冷酷触发**来训练我们的 AI。**每个玩家开始时都选择合作，但如果对手在任何回合选择背叛，那么他们将永远选择背叛**。通过这种方式，每个玩家都有动机维持合作，因为他们知道一旦他们背叛，对手将永远背叛他们。

在强化学习的环境中，可以将此作为状态 -动作 -奖励的设定。状态包括自己和对手的历史行动，动作是选择扩军或不扩军，奖励是基于博弈的结果。如果我们用深度 Q 学习等强化学习算法进行训练，智能体可能最终学习到使用类似冷酷触发的策略，以实现最大的长期收益。

这个结论对于游戏 AI 的设计是有很大帮助的。例如游戏《欧陆风云 4》等历史战略游戏，国家为了发展显然可以与其它国家形成同盟，但是如果同盟国在战争中背叛，那么这个国家就会永远记恨这个国家，从而在以后的游戏中不会与这个国家形成同盟。

这个结论在多智能体中会有更多的应用。**在多智能体中，我们可以使用冷酷触发来训练我们的 AI，使得我们的 AI 具有道德性、合作性。** 我会在多智能体的部分再来探讨这一点。

3 生成对抗网络的贝叶斯博弈模型

《人工智能基础》课程也为我们讲述了生成对抗网络 GAN 的具体实现。**为什么 GAN 能取得很好的效果？因为这是一个对抗性攻击和防御的博弈过程。** 在对抗性机器学习中，攻击者和防御者之间的博弈是核心问题。攻击者试图通过添加精心设计的扰动来欺骗神经网络，而防御者则试图使网络对这些攻击具有鲁棒性。这个过程可以被建模为一种博弈，通过博弈论可以寻找到最佳的攻击和防御策略。

事实上，生成对抗网络可以被视为**一种特殊的贝叶斯博弈模型**。生成器 G 和判别器 D 分别尝试去拟合这个博弈的贝叶斯均衡。我们尝试用信息不完善的无限重复动态博弈来表示出来这个贝叶斯博弈的模型：

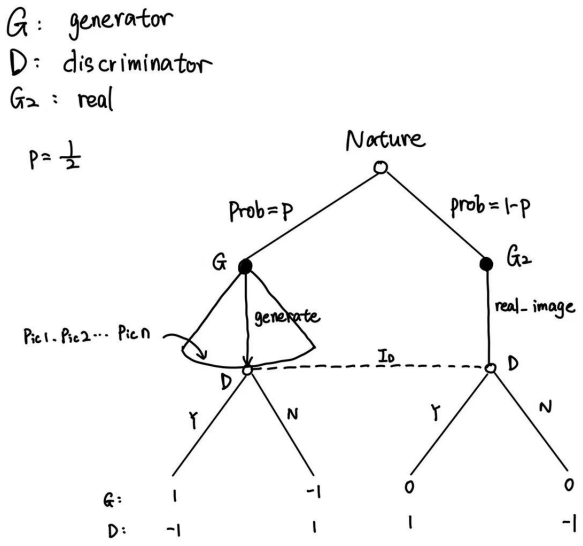

图 1: 博弈树

在这个博弈树中，我们可以看到，生成器和判别器是两个博弈者，生成器的策略是生成图片 $P_1, P_2 ... P_n$，判别器的策略是判断图片是否为真。但是判别器并不知道图片是生成的还是真实的，所以判别器在一个阶段的博弈中只有一个信息集。我们可以用收益矩阵来表示这个博弈：

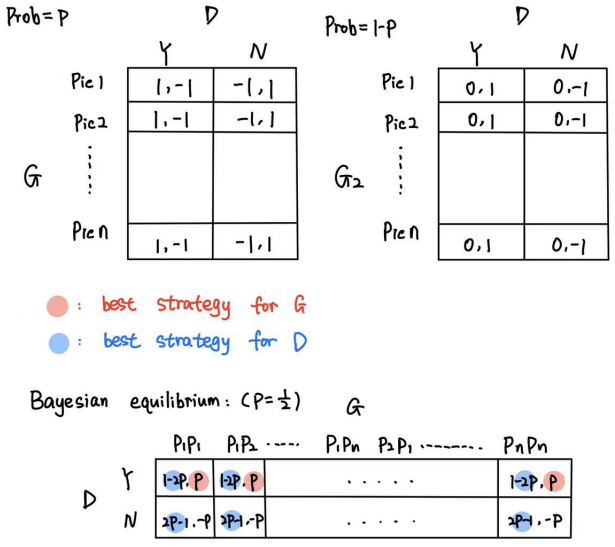

图 2: 收益矩阵

根据贝叶斯博弈，我们**可以很好的解释为什么在 GAN 网络的训练中，我们一般设置真实图片和生成器生成图片的比例为 1:1。**

用抽象的语言解释，设置为 1:1 的理由在于，这种均衡的比例可以确保在训练过程中，判别

器对真实样本和假样本的接触机会是平等的，从而防止它倾向于过度专注于真实样本或假样本。如果真实样本过多，判别器可能过度倾向于识别真实样本的特征，而忽略假样本的特征，反之亦然。

利用收益矩阵的可视化，我们可以发现，如果 $p = \frac{1}{2}$，那么判别器的最优策略是 Yes 和 No 的混合策略，否则则会倾向于使用 Yes 或 No。

同时我们发现，在达到贝叶斯均衡时，判别器使用的策略 Y 满足收益 $U_y = 1 - 2p$，而使用策略 N 的收益为 $U_n = 2p - 1$。对于生成器来说，我们很难保证生成器不会在收益相同的时候选择不处于均衡的 N，生成器反而可能以 1:1 的比例使用 Y 和 N 策略，这也就解释了为什么在训练生成对抗网络的最后，生成器会尝试生成能使 D 使用 Y 的图片，而 D 会在训练末期产生 1:1 的策略比。

4 多智能体

4.1 多智能体强化学习：子博弈完善均衡

在强化学习的部分，我已经讨论了使用冷酷触发策略的 AI 可以通过强化学习的方法变得有道德性、合作性。但是，冷酷触发的机制在博弈论中是相对于双方使用的，如果只有一方使用冷酷触发，那么另一方可以根据使用冷酷触发的博弈者的策略，针对性的使用另一套策略，这会导致整个博弈构不成一个稳定的状态。所以我们使用冷酷触发训练的 AI 应该还有一个前提，就是训练是处在一个多智能体的环境中的。

在单智能体的强化学习中，虽然只有一个模型，但是在训练过程中自己会和自己进行博弈，在某种程度上也可以实现训练。但是博弈论更多的则是运用在多智能体环境的训练中。

由于训练的轮数很长，且较大模型的学习率一般不会很高，**我们可以将训练的过程视为一个无限重复博弈的过程**。于是，我们可以利用博弈论的知识，设计更多、更巧妙的冷酷触发机制，来实现更多种类的 AI。原则上，只要满足策略的收益严格大于单一阶段单纯策略的纳什均衡收益，就可以使用冷酷触发来获得相应的策略 [3]。如果满足策略的收益严格大于单一阶段的 Minmax 收益，那么就可以使用温和、有限的惩罚来获取相应的策略。这体现了多智能体的一个优越性，即 **多智能体环境下训练 AI 较单智能体环境可以满足更多的要求**。

4.2 多智能体分散式学习：弊端

在上述多智能体强化学习的过程中，我们假设不同的模型是不同的博弈者，使用不同的策略进行训练，这种训练是多智能体的分散式学习。多智能体分散式学习是指在一个多智能体环境中，每个智能体都各自进行学习，而不是集中式的进行学习。这种学习方式的主要特征是独立性和分散性。每个智能体都会基于自己的观测和经验独立地更新自己的策略。相对于简单的单智能体训练，或者多智能体的集中性训练，多智能体的分散式学习具有明显的优势和劣势，而多智能体分散式学习的优点和缺点也是可以通过博弈论来进行解释的。在这里，我们先只关注多智能体分散式学习的一个弊端，并且尝试进行改进。

由于多智能体分散式学习和博弈论中的博弈模型非常相似，所以二者的弊端也是相似的。在博弈论中存在“囚徒困境”，正是因为博弈者要求自己的利益最大化，反而没有顾及到社会的利

益，这是博弈的**外部性**的影响。在多智能体分散式学习中，每个智能体都是独立的，所以分散式学习的系统对于“整体利益最大化”的维护是一个亟待解决的问题。

我们考虑如下例子：

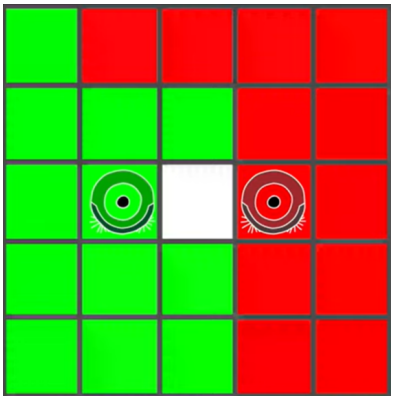

图 3: 分散式学习的困境

在上图中，两个扫地机器人在思考是否要打扫中间的部分，在第一次博弈中，出于加分的考虑，它们会拥有打扫空白区域的倾向。但是当它们全部进入格子，产生“相撞”的负得分之后，在下一次面对这种情况时，很有可能会选择不打扫中间的区域，从而导致整体达不到“清扫整个房间”的效果。

解决“无法使社会利益最大化”的问题，在博弈论中有相关的方法，即利用 VCG 机制 [4]，为每一个博弈者添加一个转移收益 t。通过这种机制来约束博弈者们的收益，我们可以得到一个全部博弈者“说真话”的均衡。

在多智能体系统中，VCG 机制可以被用于分配共享资源。假设有一个多机器人系统，其中机器人需要共享某些资源，例如电力或计算能力。每个机器人都有自己的任务，并且对资源的需求不同，这种需求对于其他机器人是私有信息。在这种情况下，可以使用 VCG 机制来决定资源的分配。每个机器人首先提交一个报价，表示它愿意为资源支付的价格。然后，根据所有机器人的报价，使用 VCG 机制分配资源，并确定每个机器人应支付的价格。

这个过程有两个关键的特性。首先，它**提供了一个诚实的机制**，意味着每个机器人的最优策略是真实地报告其对资源的评价。其次，它**实现了社会福利最大化**，也就是说，资源分配的结果是所有机器人的评价之和最大的分配。

通过这种方法，我们可以解决多智能体分散式学习中的“囚徒困境”，使得多智能体分散式学习的系统可以达到“社会利益最大化”的效果。

5 总结

根据上面的讨论，我们可以看出，博弈论在 AI 中的应用是非常广泛的。在强化学习中，我们可以使用博弈论的方法来优化强化学习的训练，使得训练的结果更加合理；在生成对抗网络中，我们可以使用博弈论的方法来解释 GAN 的训练过程，从而使得 GAN 的训练更加合理；在多智

能体系统中，我们可以使用博弈论的方法来解决多智能体系统中的囚徒困境，使得多智能体系统可以达到“社会利益最大化”的效果。

总而言之，博弈论的使用，可以是我们训练出来的 AI 具有更强大的理性。博弈论很好的量化了理性，这也给 AI 发展提供了非常好的一个便利，因为我们**可以“量化”AI 的理性**了。

让我们回到通用人工智能的研究上。AGI 是否可以通过多智能体训练以及博弈论的辅助实现？我认为是非常有可能的。早在 2006 年，日本轻小说《刀剑神域》中就有了这样的设定：科学家通过创造一个多智能体系统形成的世界，在这个世界中让 AI 模拟生存，从而让 AI 获得了自我意识。最终成为一个泛用人工智能。虽然这个设定是虚构的，但是它的思想是非常有意义的。如果我们可以让 AI 在一个多智能体系统中进行训练，并且通过奖励/惩罚机制让它变得更加理性、更加具有人性，那么我们是否也可以做到这一点呢？虽然这种方法需要经过算力、算法、空间等多个方向的挑战和质疑，但是这种思路对于多智能体环境来说显然是非常宝贵的。

6 致谢

感谢邓小铁老师、杨耀东老师及其他老师在《人工智能初级研讨班》课程中的讲解，让我对多智能体、强化学习及其他各个方面有了更深的理解。

感谢李文新老师在《人工智能基础》课程中的讲解，让我对人工智能的基础算法有了更深的理解，以及学会尝试着去实现它们。

感谢陶林老师在《社会博弈论》课程中的讲解，让我对初步认识了博弈论，以及博弈论在社会科学中的应用。

感谢《刀剑神域》制作组，为我提供了非常好的灵感，同时让我在很小的时候就对多智能体和通用人工智能产生了兴趣。

参考文献

[1] Martin J. Osborne - An Introduction to Game Theory

[2] Stuart Russell , Peter Norvig - Artificial Intelligence: A Modern Approach (4th Edition)

[3] James W. Friedman - A Non-cooperative Equilibrium for Supergames

[4] Y. Narahari - Game Theory and Mechanism Design