计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(3):15−24 [doi: 10.15888/j.cnki.csa.008973]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 TD3 算法的自动协商策略①

陈佐明, 詹捷宇

(华南师范大学 计算机学院, 广州 510631)
通信作者: 詹捷宇, E-mail: zhanjieyu@scnu.edu.cn

摘　要: 协商是人们就某些议题进行交流寻求一致协议的过程. 而自动协商旨在通过协商智能体的使用降低协商成本、提高协商效率并且优化协商结果. 近年来深度强化学习技术开始被运用于自动协商领域并取得了良好的效果, 然而依然存在智能体训练时间较长、特定协商领域依赖、协商信息利用不充分等问题. 为此, 本文提出了一种基于 TD3 深度强化学习算法的协商策略, 通过预训练降低训练过程的探索成本, 通过优化状态和动作定义提高协商策略的鲁棒性从而适应不同的协商场景, 通过多头语义神经网络和对手偏好预测模块充分利用协商的交互信息. 实验结果表明, 该策略在不同协商环境下都可以很好地完成协商任务.

关键词: 自动协商; 协商策略; 深度强化学习; TD3 算法; 偏好预测

引用格式: 陈佐明,詹捷宇.基于 TD3 算法的自动协商策略.计算机系统应用,2023,32(3):15–24. http://www.c-s-a.org.cn/1003-3254/8973.html

Automated Negotiation Strategy Based on TD3 Algorithm

CHEN Zuo-Ming, ZHAN Jie-Yu

(School of Computer Science, South China Normal University, Guangzhou 510631, China)

Abstract: Negotiation refers to the process in which people communicate with each other on certain topics to reach an agreement. Automated negotiation aims to reduce negotiation costs, improve negotiation efficiency, and optimize negotiation results by using negotiating agents. In recent years, deep reinforcement learning techniques have been applied to the field of automated negotiation with good results. However, there are still problems such as the long training time of agents, dependence on specific negotiation domains, and insufficient utilization of negotiation information. Therefore, this study proposes a negotiation strategy based on the TD3 deep reinforcement learning algorithm, which reduces the exploration cost of the training process through pre-training and improves the robustness of the negotiation strategy by optimizing the state and action definitions, so as to adapt to different negotiation scenarios. In addition, it makes full use of the interaction information of the negotiation by multi-head semantic neural network and opponent preference prediction module. The experimental results show that the strategy can perform the negotiation task well in different negotiation environments.

Key words: automated negotiation; negotiation strategy; deep reinforcement learning; TD3 algorithm; preference prediction

协商是日常生活和工作中用于解决矛盾和冲突的常用方法, 指的是人们就某些议题进行交流寻求一致协议的过程[1]. 在很多应用场景, 如电子商务中的价格商议、资源分配问题等需要采用协商的方法. 然而传统的人与人之间的协商往往依赖于参与者长时间的接触和沟通, 花费了参与者过多的精力, 而且由于人的精

① 基金项目: 国家自然科学基金青年基金 (62006085)
收稿时间: 2022-07-27; 修改时间: 2022-08-26; 采用时间: 2022-09-16; csa 在线出版时间: 2022-11-18
CNKI 网络首发时间: 2022-11-22

力有限同时容易受到情绪等因素的影响可能导致协商策略等不能发挥应有的效果, 从而影响了协商的顺利开展. 随着信息时代的到来, 学者们开始关注如何通过计算机相关技术使得协商更加自动化和智能化, 从而降低人类开展协商活动所需的成本, 提高协商效率并且优化协商结果. 为了实现上述的目标, 越来越多的协商智能体被用于自动协商中, 代替或者帮助人类进行协商并达成协议. 目前自动协商已经被成功运用于管理、经济、网络、交通、网格计算等多个领域, 包括供应链管理[2]、雇主和雇员的工资谈判[3]、Wi-Fi 信道分配[4]、物流公司车辆路线规划[5] 和网格资源分配[6] 等多种应用场景. 随着计算机科学与人工智能技术的发展以及各行业对协商效率和协商效果要求的增加, 自动协商的研究将具有更为广阔的应用意义.

协商按照不同的标准可以区分为不同的类型. 从议题数量的角度来看, 协商可以分为单议题协商和多议题协商. 从参与人数来看, 协商可以分为双边协商和多边协商. 在本文中, 我们主要考虑双边多议题协商的情况. 在基本的双边协商中, 协商协议、协商场景和协商策略共同构成了整个协商系统. 协商协议规定了参与协商的智能体应遵循的规则, 只有制定了协商规则, 协商才能顺利进行. 其中最常见的协议是交替报价协议, 即协商双方轮流出价, 直到达成协商一致或协商失败为止. 协商场景决定了双方的具体协商内容, 通常表示为 Ω. 每个协商场景都可以被视为一个结果空间. 结果空间的元素 $\omega \in \Omega$ 表示协商各方之间要交换的报价. 协商策略是协商智能体采取行动所依据的策略. 协商策略通常被设计为一个具体考虑到时间、对手行为或其他因素的函数. 时间依赖策略考虑了相对时间, 并根据超参数平滑或激进地进行折衷. 行为依赖策略则根据对手过去的行为修改自己的反应, 例如针锋相对策略.

为了更好地研究自动协商, 研究者们将协商策略分解为 3 个组成部分: 出价策略、接受策略和对手建模. 如果对对手的报价不满意, 出价策略用于选择还价, 常见的如机器学习算法[7], 启发式算法[8] 等被用于出价策略的实现. 接受策略用于确定何时接受报价并达成协议. 对手建模[9] 是指一系列用于探测对手偏好的技术. 贝叶斯学习 [10,11], 核密度估计[12], 神经网络[13,14] 和多项式插值[15] 等技术被用于了解对手的接受策略和出价策略.

随着深度强化学习技术在各领域中的成功应用, 深度强化学习算法正逐步应用于自动协商领域当中并取得了良好的效果[16–18]. 然而当前工作依然存在一些问题, 如智能体训练时间较长、训练依赖于特定领域、对协商信息的利用不够充分等, 在这项工作中, 我们提出了一种基于 TD3 深度强化学习算法的协商策略, 通过预训练降低训练过程的探索成本, 通过优化状态和动作定义提高协商策略的鲁棒性, 使其能适应于不同的协商场景. 为了更好地利用协商的交互信息, 我们采用了多头语义神经网络, 并调整状态以适应网络的输入, 同时构建了一个神经网络来预测对手偏好, 用于选择更适合对手的提议.

1 相关工作

随着机器学习研究的发展, 各种机器学习方法被使用并结合在一起, 以解决自动协商领域中的许多问题, 不断提高策略的性能. 如文献 [19] 通过贝叶斯学习估计议题对对手的重要性来学习对手的偏好, 文献 [20] 将递归神经网络和强化学习相结合来解决人机交互的物品分配协商问题. 近年来, 随着深度强化学习的迅速发展, 一些研究开始使用深度强化学习方法来解决自动协商问题, 以最大限度地提高协商过程中的总体收益, 包括模拟接受策略、将深度强化学习算法部署到电子商务环境等. 下面我们总结了在自动协商研究中应用深度强化学习技术的相关工作, 这些工作表明深度强化学习技术可以有效解决一些协商中的问题, 在方法论上对本文具有一定的借鉴意义. 同时, 我们也根据已有工作的优缺点调整了我们的研究重心和研究方法, 解决目前存在的问题.

文献 [16] 模拟了一个电子市场环境, 提出了一个基于深度强化学习的协商模型, 称为 ANEGMA. 作者提出了市场密度、供需比、协议区域和结束期限对协商有显著影响的假设, 并对此进行了证明. 但是, 该工作的实验设置仅考虑了少数几种类型的传统对手和单一问题场景, 这不足以显示模型的鲁棒性, 为了解决单一问题场景的问题, 我们使用效用值作为输出, 再根据效用值和报价生成方法得出要返回的报价. 文献 [17] 使用深度强化学习算法来构建竞价策略, 并以概率分布的形式输出每一个议题的值, 作者测试了 3 个不同的分布, 分别是正态分布、柯西分布和贝塔分布, 并评估了不同分布之间的优缺点. 然而, 该工作的缺点是其

网络结构必须随着问题的数量而变化, 这是阻碍其应用范围的一个主要问题. 文献 [21] 使用深度强化学习来构建接受策略, 该策略可以决定智能体何时接受对手的报价. 结果表明, 基于深度强化学习的接受策略在不同指标上持平或优于其他预定义的接受策略. 该工作的研究重心主要在接受策略, 而我们则更加关注对竞价策略的研究. 文献 [22] 考虑到将效用函数值作为神经网络的输出, 使得模型能适应不同的协商场景, 同时其提出了多策略的概念, 并使用对手分类器匹配最佳的应答策略, 但在选择报价时没有充分考虑对手偏好, 本文引入了对手建模模块, 从而更好地预测对手的偏好. 文献 [23] 也使用到了学习机制和策略重用机制, 针对未接触过的对手设计新的策略, 并根据对手行为选择最佳应答策略. 相比于构造多个策略的研究思路, 我们希望构建一个更加通用的策略, 并通过添加对手建模等模块, 充分探索对手的偏好. 文献 [24] 整合了语言交流能力和竞价策略, 实现了基于深度强化学习的多渠道自动协商框架. 该工作主要关注将语言表达与竞价策略相结合, 我们则更加关注竞价策略本身的设计上.

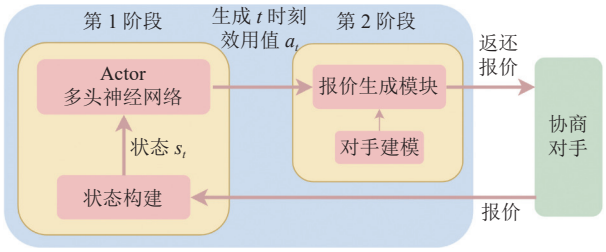


图 1 深度强化学习策略框架

2 协商策略框架

本文设计了一个基于 TD3 深度强化学习算法的协商策略框架 (如图 1 所示). 该策略由两个阶段组成: 在第 1 阶段, 协商智能体在得到对手返回的报价后, 根据协商过程中的历史序列作为当前环境的观测, 对环境状态进行构建, 将状态输入多头语义神经网络, 基于 TD3 算法输出得到对应动作, 即当前时刻效用值. 值得注意的是, 在第 1 阶段所使用的网络需要经过两个步骤的训练. 在初始状态下, 网络作为一个随机策略, 我们使用专家经验对网络进行预训练, 大大减少其在探索期间的资源开销和时间开销, 随后我们使用 TD3 算法进一步对网络进行更新, 得到最终在实验部分所使用到的网络. 在第 2 阶段, 由于我们需要给对手返回一个具体的报价, 因此我们根据在第 1 阶段得到的效用值, 结合对手建模方法和报价生成方法, 得到最终要返回的报价. 该报价综合考虑了自身的效用和对手的偏好, 能较好地完成不同场景下的协商任务, 达到协商双方双赢的结果.

2.1 模型预训练

由于深度学习被融合进强化学习算法当中, 我们得以使用监督学习的方式对 Actor 网络模块进行了预训练. 这样, 可以大大降低训练过程的探索成本. 为了预先训练策略, 我们需要解决的第 1 件事是收集合适的 Actor 网络标记数据.

由于我们对状态进行了重新定义, 我们需要自己收集模拟数据, 以匹配 DRL 模型的输入形式. 我们使用了一个协商模拟环境来收集两个协商智能体之间每轮协商会话的过程. 之后, 可以为 DRL 模型构建输入数据. 在本文中, 每一时刻的状态是由不同形式的协商轨迹组成的集合. 因此, 记录下两个协商智能体之间每轮协商的过程就能构建出相应的状态-动作对.

我们可以使用自动协商智能体竞赛 (ANAC[25]) 中的一些协商智能体收集训练数据, 本文中自身智能体使用 Parscat, 对手则被设置为 Atlas3, RandomDance 和 Parscat 等 3 种不同智能体. 在收集训练数据后, 我们使用反向传播对 Actor 网络进行预训练. 训练的超参数如表 1 所示. 在预训练 Actor 网络后, 我们通过固定 Actor 网络的形式继续预训练 Critic 网络.

表 1 预训练参数

训练参数	值
训练轮数	50
学习率	2E–4
批量大小	128
损失函数	MSE
优化器	Adam

2.2 深度强化学习方法

2.2.1 Actor-Critic 架构

Actor-Critic 架构是最重要的 RL 方法之一. 在 Actor-Critic 架构中, 两个神经网络分别扮演演员和评论家的角色. Actor 网络根据当前状态s_t确定智能体的动作. 在执行动作 a 后, 智能体能够获得由动作 a 所引起的奖励信号 r 和下一个状态s_{t+1}. Critic 网络在得知奖励为 r, 下一个状态的估计值为s_{t+1}的情况下, 可以使

用时间差分法更新其参数, 公式如下:

$$y_t = r_t + \gamma Q(s_{t+1}, \mu(s_{t+1}|\theta^{\mu})|\theta^{Q})$$

其中, θ^{μ}和θ^{Q}分别代表 Actor 网络和 Critic 网络的参数.

在策略梯度算法当中, 策略梯度的更新需要利用到整个轨迹的累计回报, 无法实现单步的策略梯度更新, 会造成方差较大的问题. Actor-Critic 架构结合了 Value-based 和 Policy-based 两类算法, Actor 能在连续动作空间中轻松地得到输出, Critic 则能够预测当前状态下所能获得的奖励, 使得 Actor 能够实现单步更新, 解决了维度爆炸和更新效率低等问题.

2.2.2 TD3 算法

为了实现让策略在少数场景下的训练能适应不同的协商场景, 相比于输出具体的报价, 我们使用效用值作为 Actor 网络的输出. 而由于输出的效用值是连续型的, 因此, 协商策略的制定需要处理连续动作空间, 并根据当前环境状态做出具体动作, 适合使用 Actor-Critic 架构中的确定性策略梯度算法[26], 该算法能较好地处理连续动作空间的问题, 并保证较好的性能. 因此, 我们选用双延迟深度确定性策略梯度算法 (TD3) 作为主要算法, 并通过增加预训练, 调整网络结构等方法, 使其能更好地适应当前的协商问题.

TD3 算法[27] 是深度确定性策略梯度的改进版本, 其仍然属于 Actor-Critic 架构的算法, Actor 用于提供给定观测状态的具体动作值, Critic 则用于评估 Actor 作出的动作的好坏. 它针对原始算法进行了一些改进, 技术要点包括使用两个网络近似 Q 函数、延迟策略更新和目标策略平滑正则化等. 具体如下.

(1) 由于函数的近似误差可能导致高估其真实值, TD3 算法通过分别学习 2 个目标 Q 函数, 并取两个函数输出的较小值, 可以在一定程度上缓解 Critic 的过高估计问题.

(2) 延迟策略更新用于使算法更加稳定, 训练效率更高. 在算法训练过程中, Actor 和 Critic 是同时在进行更新的, Critic 的变化会导致最优策略的变化, 而 Actor 的更新是由 Critic 网络的输出所决定的, 因此 Critic 网络频繁的变化会使得 Actor 网络的更新变得不稳定, 甚至导致无法求得最优解. 为了解决这个问题, TD3 算法引入了延迟策略更新的方法, 其主要思路是使得 Critic 网络的更新频率大于 Actor 网络的更新频率, 缓解了 Critic 网络的频繁变化导致 Actor 网络的不稳定.

(3) 目标策略平滑正则化作为一种正则化方法, 其基于这样一种假设: 相似的动作有着相似的值. 通过在目标策略上添加噪声, 缓解估值函数的过拟合现象, 能让估值函数更加平滑.

这些改进缓解了 Q 函数值的高估问题, 让 Actor 网络的训练更加稳定, 增强了算法的稳定性. 具体的算法流程如算法 1 所示.

算法1. TD3算法流程

初始化经验回放缓冲区R, 大小为100 000
初始化Q网络$Q_1(s,a|\theta^{Q_1})$、$Q_2(s,a|\theta^{Q_2})$和策略网络$\mu(s|\theta^{\mu})$, 参数分别为θ^{Q_1}、θ^{Q_2}和θ^{μ}
初始化目标网络Q_1'、Q_2'和μ', 参数分别为$\theta^{Q_1'}\leftarrow\theta^{Q_1}$、$\theta^{Q_2'}\leftarrow\theta^{Q_2}$和$\theta^{\mu'}\leftarrow\theta^{\mu}$
1) for episode =1:M do
2) 初始化环境状态s_1
3) for t =1:T do
4) 根据策略网络和高斯噪声生成当前时刻动作 a_t
5) 执行动作a_t, 得到奖励r和下一时刻环境状态s_{t+1}
6) 将过去一时刻状态转移过程(s_t,a_t,r_t,s_{t+1})存储到经验回放缓冲区中
7) 从经验回放缓冲区中采样小批量的状态转移过程(s_t,a_t,r_t,s_{t+1})
8) for 每一个状态转移过程 do
9) 使用目标策略网络得到状态s_{t+1}下的动作a_{t+1}
10) 得到目标值 $y_t=r+\gamma\cdot\min\left(Q_1'(s_{t+1},a_{t+1}+\varepsilon),Q_2'(s_{t+1},a_{t+1}+\varepsilon)\right)$
11) 使用如下公式和Adam优化器更新Q网络参数:
12) $\theta^{Q_i}\leftarrow\min(y_t-Q_i(s,a))$
13) if t满足策略网络更新条件 then
14) 使用策略梯度更新策略网络:
15) $\nabla_{\theta^{\mu}}J(\theta^{\mu})=\nabla_a Q_1(s,a)|_{a=\pi_{\theta^{\mu}}(s)}\nabla_{\theta^{\mu}}\pi_{\theta^{\mu}}(s)$
16) 使用如下公式更新目标网络
17) $\theta^{Q_i'}\leftarrow\tau\theta^{Q_i}+(1-\tau)\theta^{Q_i'}$
18) $\theta^{\mu'}\leftarrow\tau\theta^{\mu}+(1-\tau)\theta^{\mu'}$
19) endif
20) endfor
21) endfor
22) endfor

2.2.3 状态空间与动作空间的定义

在使用深度强化学习解决具体问题前, 我们需要对算法的状态, 动作和奖励函数的具体定义作出规定, 状态空间必须包含协商过程中所需要的全部信息. 在以往的一些工作中, 对手的出价被视为当前状态, 并以具体还价的形式采取行动. 使用这种形式的状态和动作, 深度强化学习方法将被局限于某一具体协商场景, 并失去算法的适用性. 为此, 文献 [22] 提取出协商过程中共有的属性 (效用值) 作为动作的输出, 并根据效用值选取离该效用值最接近的报价. 但其在生成报价过程中没有考虑到对手的偏好, 仅根据自身的利益作出了选择. 为了解决这个问题, 我们加入了对手建模的方法用于探测对手的偏好, 并重新设计了一个新的状态

定义, 即 3 种不同形式的效用序列的组合, 充分利用协商过程中所产生的可用信息.

为了实现在不同领域进行协商的可能性, 动作空间被定义为输出下一个时间步的效用值, 得到效用值后再根据效用值生成相应的具体报价. 状态和动作被定义为:

$$\begin{cases} s_t = \{s_{t1}, s_{t2}, s_{t3}\} \\ s_{t1} = \left\{t_r, U_s\left(\omega_s^{t-6}\right), U_s\left(\omega_s^{t-5}\right), U_s\left(\omega_s^{t-4}\right), \right. \\ \quad \left. U_s\left(\omega_s^{t-3}\right), U_s\left(\omega_s^{t-2}\right), U_s\left(\omega_s^{t-1}\right)\right\} \\ s_{t2} = \left\{t_r, U_s\left(\omega_o^{t-6}\right), U_s\left(\omega_o^{t-5}\right), U_s\left(\omega_o^{t-4}\right), \right. \\ \quad \left. U_s\left(\omega_o^{t-3}\right), U_s\left(\omega_o^{t-2}\right), U_s\left(\omega_o^{t-1}\right)\right\} \\ s_{t3} = \left\{t_r, U_s\left(\omega_s^{t-3}\right), U_s\left(\omega_o^{t-3}\right), U_s\left(\omega_s^{t-2}\right), \right. \\ \quad \left. U_s\left(\omega_o^{t-2}\right), U_s\left(\omega_s^{t-1}\right), U_s\left(\omega_o^{t-1}\right)\right\} \\ a_t = u_s^{t+1} \end{cases}$$

其中, t_r表示相对时间, $U_s(\omega)$表示报价ω的效用值. ω_s^t和ω_o^t表示自己和对手在时间 t 提出的报价. 为了更好地模拟现实协商场景, 我们设定无法直接获得对手具体的偏好信息. 因此, 在映射不同的报价到效用值的过程中, 仅使用我们自身的效用函数, 3 个序列都是使用自身的效用函数实现从报价到效用值之间的映射. 考虑到无法得知对手的具体偏好, 我们充分利用协商过程中的历史数据, 形成组织形式不同的 3 个子序列, 从而更加充分地利用已有的信息实现更精确的预测. 第 1 个子序列是自身智能体在过去 6 步的报价, 第 2 个子序列是对手在过去 6 步的报价, 第 3 个子序列则是双方过去 3 步的交替报价过程.

时间 t 的动作a_t是我们想要提出的报价的效用值. 通过这种方式, 智能体可以适应不同的协商领域, 而不考虑提出具体的报价. 然而, 为了协商成功, 必须将效用价值转化为具体报价. 具体方法在第 2.4 节中介绍.

2.2.4 奖励函数的定义

为了使深度强化学习策略不断进步, 我们需要仔细设计奖励函数. 在协商场景中, 我们鼓励智能体达成协议, 同时惩罚失败的例子. 在达成协议后, 会给予一个正向的奖励值, 该值介于 0 到 1 之间, 其含义表示为达成协议的报价所对应的效用值. 当协商破裂后, 设置奖励值为−1, 即不鼓励该行为的出现. 在协商过程中的某些特殊情况下, 考虑到前期智能体在探索阶段会达成效用值较低的协议, 因此即使是效用值较低的协议也应受到惩罚. 其余情况下, 奖励值则设置为 0. 奖励函数 R 的具体定义如下:

$$R(s_t, a_t, s_{t+1}) = \begin{cases} U_s(\omega_a), & \text{达到协议}\omega_a\text{且效用值不小于}u_r \\ -(1-U_s(\omega_a)), & \text{达到协议}\omega_a\text{且效用值小于}u_r \\ -1, & \text{达不成协议} \\ 0, & \text{其他情况} \end{cases}$$

其中, ω_a表示达成协议的报价而u_r表示智能体能接受的最低效用值. 若达成协议但效用值低于该值, 则通过奖励函数修正并减少该情况的出现.

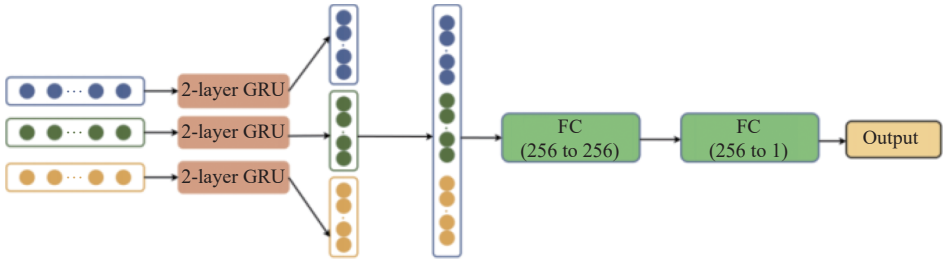
2.2.5 Actor 与 Critic 的网络结构

虽然我们的 TD3 策略作为投标策略包含 6 个神经网络, 但它主要由两种类型的网络组成: Actor 网络和 Critic 网络. Actor 网络被作为算法中的决策者, 并根据环境的当前状态作出动作. Actor 网络结构如图 2(a) 所示. 为了有效利用协商信息, 我们提出了多头输入语义网络的网络结构, 并重构智能体观测的状态以适应新的网络结构. 状态是协商轨迹的不同组织形式, 以便更好地利用有关协商的不同语义信息. 对于时间序列形式的状态输入, 我们使用 GRU 模块作为输入信息的特征提取模块. 在经过 GRU 模块后, 由于多头输入的设计, 我们需要连接不同通道的特征向量. 经过拼接后, 特征向量将通过一个多层感知机模块, 最后输出我们需要的结果.

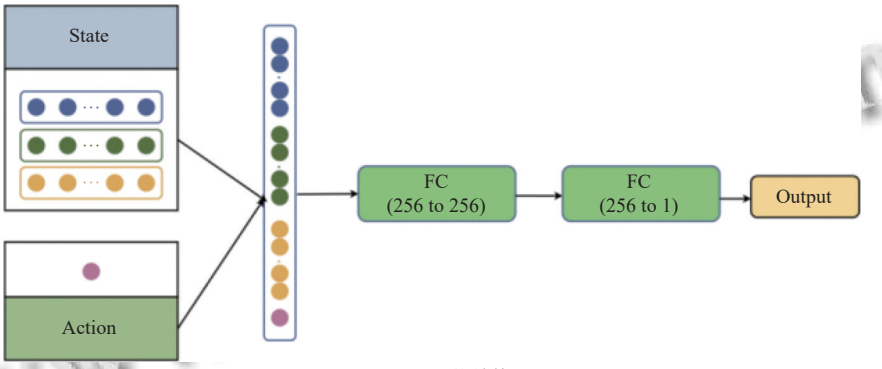
Critic 网络将状态和行为作为输入, 其目的是引导 Actor 网络做出更好的决策. 与 Actor 网络的 3 个向量的输入相比, Critic 包含 4 个向量作为输入, 包括动作向量和由 3 个向量组成的当前状态. Critic 网络结构如图 2(b) 所示. 在 TD3 算法中, 为了缓解 Actor-Critic 算法的高估问题, 使用了两个 Critic 网络来构建价值估计. 在获得两个值后, 取较小的值作为最终估计, 以减少偏差, 从而减轻高估问题.

2.3 对手建模方法

如前文所提到, 我们使用深度强化学习策略充当竞价策略, 并输出下一时间步的效用值. 在得到效用值以后, 基于该效用值得到具体报价是必不可少的. 然而, 对于自我效用一致的不同报价, 对于对手来说是具有不同偏好的. 如何在已知效用值的情况下, 判断出对手更偏好哪一个报价, 对于更好地达成协商一致具有积极的正向作用. 为了更好地探索对手的偏好, 我们提出了一种基于历史协商序列数据和神经网络的对手建模方法, 用于评估对手对不同报价的偏好. 具体的网络结构如图 3 所示. 输入向量是对手返回报价所对应的向量, 输出则是对手关于该报价的效用的预测值.



(a) Actor 网络结构



(b) Critic 网络结构

图 2 Actor 和 Critic 网络结构

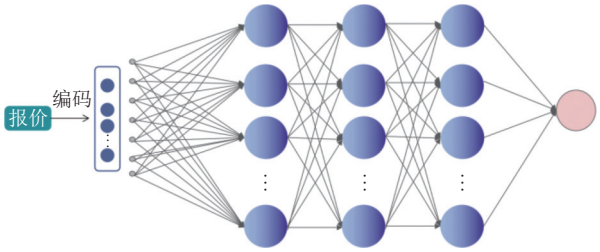


图 3 对手建模网络结构

该方法可归类为回归问题, 输入是由报价经过编码所生成的数值向量, 输出则是对手对该报价的偏好预测值. 训练数据的获取则是基于这样的假设: 在一轮协商当中, 对手的报价顺序整体呈现效用下降的趋势. 在得到训练数据后, 使用梯度反向传播对神经网络进行训练.

在协商对手及协商场景切换的情况下, 需要根据具体报价的议题数量对神经网络进行重新初始化, 以适应不同议题数量的协商场景. 在和对手协商过程中, 不断依赖历史协商序列进行权重的更新.

2.4 报价生成方法

在报价生成过程中, 并不仅取决于协商智能体所提供的效用值, 同时还考虑到对手对当前协商场景的偏好情况, 使用到第 2.3 节中的方法, 对对手偏好进行估计, 进而生成报价. 首先, 我们得到该策略在每个时间步长上给出的效用值 U. 然后, 我们根据 U 值对其附近范围内的报价空间进行 10 次采样得到 10 个不同的报价, 在本文中范围设置为 (U–0.05, U+0.05). 经过采样过后, 通过使用对手建模中的神经网络估计每个报价的效用值. 基于神经网络对不同报价的估计, 最终得出最适合对手的报价. 报价生成方法如算法 2 所示.

算法2. 报价生成方法

输入: Actor 网络在 t 时刻生成的效用值 U

输出: 报价ω_t

```
1) for episode =1:M do
2)    初始化报价池O
3)    for t =1:T do
4)       设置变量MAX=–1
5)       设置变量MAX_INDEX=0
6)       在(U–0.05, U+0.05)范围内采样n个报价并存储进报价池O中
7)       for i= 1:n do
8)          计算对手对每一个报价的效用值估计
9)          output = WeightNet(Oi)
10)         if output > MAX then
11)            set MAX = output, MAX_INDEX = i
12)         endif
13)      endfor
14)      返回ωt=O_MAX_INDEX
15)   endfor
16) endfor
```

实现对手建模的意义在于, 对于我们自己而言, 具有相同效用的多个报价对对手具有高或低效用值. 我们希望实现双赢, 给出对手喜欢的报价, 而不是仅最大化自身的利益而忽视对方的利益. 此外, 神经网络模型是一个实时模型, 它将在域变化时初始化. 模型输入层的初始化取决于域的问题编号. 在每次协商之后, 该模型使用协商历史轨迹更新其网络参数.

3 实验分析

3.1 实验设置

在本文的实验中, 使用到 Rubenstein 讨价还价协议[28] 和 ANAC 比赛中用到的用户偏好文件. 由于我们的代码是基于 Python 实现的, 因此我们使用协商多智能体系统 (NegMAS[29]) 平台进行了实验, 以更好地与 GENIUS 协商平台和 ANAC 智能体兼容. 双方的效用函数皆采用线性加权和函数的形式.

由于对深度强化学习策略的重新设计, 我们没有必要为每个协商场景重新训练模型. 这个特性大大简化了我们的训练过程, 我们只需要在几个协商场景下训练模型, 然后在多个协商场景下进行测试. 我们选择了 3 个协商场景来训练模型, 并在 27 个协商场景中评估了模型的性能. 值得注意的是, 深度强化学习策略从未在实验测试阶段所用到的协商场景上进行训练, 这在一定程度上证明了智能体的鲁棒性.

为了评估我们的深度强化学习策略的性能, 我们使用了几个指标来与其他协商策略进行比较: 平均自我效用、平均对手效用和平均社会福利. 每个指标由每个特定场景的自身和对手策略之间 30 次协商的平均结果计算得出, 这减少了协商过程中不稳定因素的影响.

在本节中, 我们提出了几种不同的实验环境, 并使用上述指标来评估模型的性能. 对于我们要评估的多个策略, 第 1 个实验是在采用不同协商策略的协商智能体之间互相进行的协商实验, 用于评估策略的一般性能. 在第 2 个实验中, 我们构建了两个新的对手环境, 即强硬环境和友好环境, 用于测试不同的协商策略在极端环境下的表现. 最后一个实验是消融实验, 用于判断对手建模模块对整个深度强化学习策略所作的贡献.

对于协商场景的选择, 我们使用 2015 年和 2013 年 ANAC 的协商场景和对应的偏好配置文件测试了该模型, 具体如表 2 所示. 为了确保对手的随机性, 在每个场景下随机选择对手的偏好配置文件, 但确保自己和对手的偏好配置文件不完全一致. 我们所选用的对比协商策略主要包括基于时间的策略和基于对手行为的策略. 在时间依赖策略当中, 根据具体的参数不同可以将其分为 Boulware, Linear 和 Conceder 这 3 种策略, 基于行为的策略会根据对手过去的报价情况做出相应的调整, 主要分为 Nice TFT 和 Naive TFT 两种策略. 实验结果表明, 与其他智能体相比, 我们设计的模型在不同极端环境下具有更好的适应性和鲁棒性.

表 2 协商场景

协商场景	结果空间	协商场景	结果空间	协商场景	结果空间
university	2 250	acquisition	384	kitchen	15 625
dinner	1 200	animal	1 152	laptop	27
holiday	1 024	camera	3 600	lunch	3 840
politics	23 040	coffee	112	niceordie	3
bank_robbery	18	defensive charms	36	outfit	128
zoning_plan	448	dog choosing	270	planes	27
car	240	fifty fifty	11	smartphone	12 000
tram	972	house keeping	384	ultimatum	9
movie	4	icecream	720	wholesaler	56 700

3.2 协商对比实验

在本实验中, 对于给定智能体列表中的每个智能体, 我们将其与其余智能体逐个进行协商实验. 具体设置如下.

自我效用计算公式:

$$U_{s\_a} = \frac{1}{(|A|-1)\cdot|D|}\sum_{d}^{D}\sum_{b}^{A/a}U_{s}^{a\_b\_d}$$

其中, a 表示自身智能体而 b 表示对手智能体, A 表示整个智能体列表, D 表示所有的协商场景列表.

对手效用计算公式:

$$U_{o\_a} = \frac{1}{(|A|-1)\cdot|D|}\sum_{d}^{D}\sum_{b}^{A/a}U_{o}^{a\_b\_d}$$

其中, a 表示自身智能体而 b 表示对手智能体.

社会福利计算公式:

$$U_{w\_a} = U_{s\_a} + U_{o\_a}$$

在给定的智能体列表中, 与第 3.3 节中的极端环境相比, 对手的类型更加均衡. 实验结果如图 4 所示. 我们从 3 个角度来判断智能体的性能: 自我效用、对手效用和社会福利. 所有指标均为通过 30 轮协商获得的平均值, 以此减少协商过程中的不稳定性带来的影响. 当使用自我效用进行比较时, 我们可以清楚地看到, 我们提出的智能体性能排名前列, 其中与 Boulware 性能相当, 相比其他智能体高出 2% 到 57.8%. 而使用对手效用对比时, 提出的智能体性能达到最优, 相比其他智能体高出 3.8% 到 37%. 与强硬的和友好的智能体相比, 提出的智能体在确保自我效用的同时, 还考虑了对手效用以实现双赢, 可以看到在社会福利指标, 该智能体仍能达到较高的效用, 与其他智能体相比增长了 11.3% 至 25.3%. 因此, 我们提出的策略相比基于时间和行为的策略在多个指标方面取得了优势, 反映了模型的稳健性, 实现了较好的利弊权衡以最终达到双赢.

3.3 强硬和友好环境对比实验

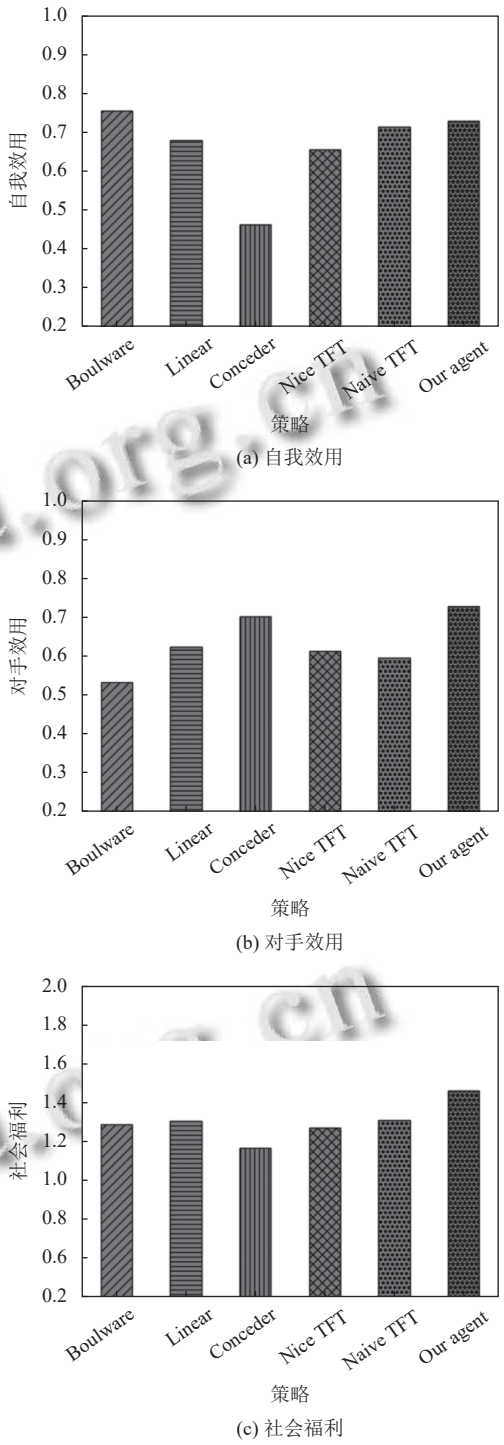
在该实验中, 我们构建了两个完全极端的环境, 即完全强硬环境和完全友好环境. 这样, 我们可以得出不同智能体在极端情况下的性能. 强硬环境中的对手采取的策略不易于妥协, 而友好环境中的对手倾向于在协商过程中会作出较大让步. 对于强硬环境, 我们选择 Boulware 策略智能体作为其基本成员. 对于友好的环境, 选择 Conceder 策略智能体作为其基本成员.

强硬环境的实验结果如表 3 所示. 在自我效用和社会福利两项评价指标中均达到最高水平. 对于自我效用来说, 我们领先其他智能体商 0.5% 至 67.3%. 另一方面, 该模型在保证自身利益的同时最大化了对手效用. 从社会福利来看, 深度强化学习模型获得了最高的社会福利效益. 表 3 所示的结果表明, 即使在极端恶劣的环境中, 该模型仍然可以实现双方的双赢.

此外, 友好环境的实验结果如表 4 所示. 在实验结果中, 友好环境中的对手更容易对对方妥协, 因此自我效用会较高而对手效用会较低. 从结果中可以看出, 我们提出的智能体具有较高的自我效用和相对较高的对手效用. 同时, 该智能体获得了最高的社会福利, 既兼顾了自我效用, 同时也考虑到双方的共同利益.

3.4 对手建模效果实验

在本节中, 为了探索策略框架中的组件的作用, 我们对对手建模模块进行消融实验, 以表明该模块对整个协商策略的贡献.



(a) 自我效用

(b) 对手效用

(c) 社会福利

图 4 协商对比实验结果

表 3 强硬环境下实验结果

评价指标	Boulware	Linear	Conceder	Nice TFT	Naive TFT	Our agent
自我效用	0.74	0.62	0.45	0.65	0.71	**0.75**
对手效用	0.65	**0.88**	0.84	0.68	0.78	0.76
社会福利	1.39	1.50	1.29	1.33	1.49	**1.51**

表 4 友好环境下实验结果

评价指标	Boulware	Linear	Conceder	Nice TFT	Naive TFT	Our agent
自我效用	0.76	0.62	0.54	0.75	0.71	**0.79**
对手效用	0.64	0.87	**0.89**	0.68	0.74	0.71
社会福利	1.40	1.49	1.43	1.43	1.45	**1.50**

对手建模模块在策略框架的第 2 阶段, 用于辅助报价生成方法, 将策略框架第 1 阶段输出的效用值转换为特定报价, 这有助于检测对手的偏好并达成双赢的局面. 为了更好地比较对手建模模块对策略的影响, 本节在基于第 3.2 节的实验环境下, 通过去除对手建模模块, 来验证该模块对整体策略的实际影响. 与第 3.2 节一样, 我们通过与给定的智能体列表中的智能体逐一进行协商实验, 该实验基于 3 个评价指标来评价智能体的好坏, 分别是: 自我效用, 对手效用和社会福利, 3 个评价指标的结果均为通过 30 轮协商获得的平均值.

表 5 给出了保留和去掉对手建模模块的模型的结果, 第 1 行记录了保留模块的实验结果, 第 2 行则记录了去除该模块的实验结果. 从表 5 可以看到, 在我们删除该模块后, 3 个评估指标都有所下降, 相比于去除该模块, 保留对手建模模块后 3 个评价指标均有所上涨, 在自我效用指标上, 提升了 4.3%, 在对手效用指标上, 提升了 14%. 可以看到, 对手建模模块对于对手效用的提升更加明显. 在社会福利指标上, 提升了 9%. 实验结果表明, 该模块在协商策略框架中起到了积极作用.

表 5 对手建模对实验结果的影响

是否保留	自我效用	对手效用	社会福利
保留对手建模模块	0.73	0.73	1.46
去除对手建模模块	0.70	0.64	1.34

4 结论与展望

本文提出了一种新的基于深度强化学习的自动协商策略框架, 将 TD3 算法应用于协商任务当中, 并修改了网络的具体实现使其更加适应当前的协商场景. 在网络的具体实现上, 我们使用了基于循环神经网络模块的多头语义神经网络. 在状态的设置中, 我们根据历史数据组织形式的不同, 划分为了 3 个子序列, 每一个子序列会具有不同的语义信息. 状态中的每一个子序列会各自经过一个 GRU 循环神经网络块, 得到的特征向量经过拼接后再进行前向传播得到输出结果. 为了更好地加快模型训练, 我们使用监督学习对模型进行预训练. 训练数据由 ANAC 智能体互动生成. 预训练加快了模型的训练速度并减少了探索时间. 为了让协商智能体获得更好的性能, 我们细化了奖励函数的设计并加入对手建模模块, 使其能更好地探测对手的偏好, 从而达成更高的效用. 结果表明, 本文提出的协商策略能够适应自动协商的任务, 并且在与各类对手协商时表现良好, 实现双赢的局面.

在未来的工作中, 我们考虑采用更多的对手建模技术, 得以充分挖掘对手的偏好, 以实现更好的双赢的协商. 此外, 我们考虑将深度强化学习用于人机交互式的协商任务, 不仅仅考虑协商的具体议题, 同时将对手的心理状态等加以考虑.

参考文献

1 Jennings NR, Faratin P, Lomuscio AR, *et al*. Automated negotiation: Prospects, methods and challenges. Group Decision and Negotiation, 2001, 10(2): 199–215. [doi: 10.1023/A:1008746126376]

2 Moghadam FS, Zarandi MHF. Mitigating bullwhip effect in an agent-based supply chain through a fuzzy reverse ultimatum game negotiation module. Applied Soft Computing, 2022, 116: 108278. [doi: 10.1016/j.asoc.2021.108278]

3 Oderanti FO, Li F, De Wilde P. Application of strategic fuzzy games to wage increase negotiation and decision problems. Expert Systems with Applications, 2012, 39(12): 11103–11114. [doi: 10.1016/j.eswa.2012.03.060]

4 De La Hoz E, Marsá-Maestre I, Giménez-Guzmán JM, *et al*. Multi-agent nonlinear negotiation for Wi-Fi channel assignment. Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems. São Paulo: International Foundation for Autonomous Agents and Multiagent Systems, 2017. 1035–1043.

5 De Jonge D, Bistaffa F, Levy J. A heuristic algorithm for multi-agent vehicle routing with automated negotiation. Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems. International Foundation for Autonomous Agents and Multiagent Systems, 2021. 404–412.

6 Haberland V, Miles S, Luck M. Negotiation strategy for continuous long-term tasks in a grid environment. Autonomous Agents and Multi-agent Systems, 2017, 31(1): 130–150. [doi: 10.1007/s10458-015-9316-2]

7 程昱, 高济, 古华茂, 等. 基于机器学习的自动协商决策模型. 软件学报, 2009, 20(8): 2160–2169.

8 党圣洁, 曹慕昆. 基于改进 NSGA-III的多边多属性自动谈判模型研究. 第十六届 (2021) 中国管理学年会论文集. 中国管理现代化研究会, 2021. 290–296.

9 Baarslag T, Hendrikx MJC, Hindriks KV, *et al*. Learning about the opponent in automated bilateral negotiation: A comprehensive survey of opponent modeling techniques. Autonomous Agents and Multi-agent Systems, 2016, 30(5): 849–898. [doi: 10.1007/s10458-015-9309-1]

10 Zhang JH, Ren FH, Zhang MJ. Bayesian-based preference prediction in bilateral multi-issue negotiation between intelligent agents. Knowledge-based Systems, 2015, 84: 108–120. [doi: 10.1016/j.knosys.2015.04.006]

11 Ji SJ, Zhang CJ, Sim KM, *et al*. A one-shot bargaining strategy for dealing with multifarious opponents. Applied Intelligence, 2014, 40(4): 557–574. [doi: 10.1007/s10489-013-0497-6]

12 Farag GM, AbdelRahman SES, Bahgat R, *et al*. Towards KDE mining approach for multi-agent negotiation. Proceedings of the 7th International Conference on Informatics and Systems. Cairo: IEEE, 2010. 1–7.

13 Carbonneau R, Kersten GE, Vahidov R. Predicting opponent's moves in electronic negotiations using neural networks. Expert Systems with Applications, 2008, 34(2): 1266–1273. [doi: 10.1016/j.eswa.2006.12.027]

14 Papaioannou I, Roussaki I, Anagnostou M. A survey on neural networks in automated negotiations. Encyclopedia of Artificial Intelligence. Hershey: IGI Global, 2009. 1524–1529.

15 Papaioannou I, Roussaki I, Anagnostou M. Multi-modal opponent behaviour prognosis in e-negotiations. Proceedings of the 11th International Work-conference on Artificial Neural Networks. Torremolinos-Málaga: Springer, 2011. 113–123.

16 Bagga P, Paoletti N, Alrayes B, *et al*. A deep reinforcement learning approach to concurrent bilateral negotiation. Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama: IJCAI, 2020. 297–303.

17 Chang HCH. Multi-issue negotiation with deep reinforcement learning. Knowledge-based Systems, 2021, 211: 106544. [doi: 10.1016/j.knosys.2020.106544]

18 Montazeri M, Kebriaei H, Araabi BN. Learning Pareto optimal solution of a multi-attribute bilateral negotiation using deep reinforcement. Electronic Commerce Research and Applications, 2020, 43: 100987. [doi: 10.1016/j.elerap.2020.100987]

19 Hindriks K, Tykhonov D. Opponent modelling in automated multi-issue negotiation using Bayesian learning. Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems. Estoril: International Foundation for Autonomous Agents and Multiagent Systems, 2008. 331–338.

20 Lewis M, Yarats D, Dauphin Y, *et al*. Deal or no deal? End-to-end learning of negotiation dialogues. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017. 2443–2453.

21 Razeghi Y, Yavuz COB, Aydoğan R. Deep reinforcement learning for acceptance strategy in bilateral negotiations. Turkish Journal of Electrical Engineering and Computer Sciences, 2020, 28(4): 1824–1840. [doi: 10.3906/elk-1907-215]

22 Sengupta A, Mohammad Y, Nakadai S. An autonomous negotiating agent framework with reinforcement learning based strategies and adaptive strategy switching mechanism. Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems. International Foundation for Autonomous Agents and Multiagent Systems, 2021. 1163–1172.

23 Wu LL, Chen SQ, Gao XY, *et al*. Detecting and learning against unknown opponents for automated negotiations. Proceedings of the 18th Pacific Rim International Conference on Artificial Intelligence. Hanoi: Springer, 2021. 17–31.

24 Gao XY, Chen SQ, Zheng Y, *et al*. A deep reinforcement learning-based agent for negotiation with multiple communication channels. Proceedings of the 33rd IEEE International Conference on Tools with Artificial Intelligence. Washington: IEEE, 2021. 868–872.

25 Jonker CM, Aydogan R, Baarslag T, *et al*. Automated negotiating agents competition (ANAC). Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco: AAAI Press, 2017. 5070–5072.

26 Lillicrap TP, Hunt JJ, Pritzel A, *et al*. Continuous control with deep reinforcement learning. Proceedings of the 4th International Conference on Learning Representations. San Juan: ICLR, 2016.

27 Fujimoto S, Van Hoof H, Meger D. Addressing function approximation error in Actor-Critic methods. Proceedings of the 35th International Conference on Machine Learning. Stockholm: ICML, 2018. 1582–1591.

28 Rubinstein A. Perfect equilibrium in a bargaining model. The Econometric Society, 1982, 50(1): 97–109. [doi: 10.2307/1912531]

29 Mohammad Y, Nakadai S, Greenwald A. NegMAS: A platform for automated negotiations. Proceedings of the 23rd International Conference on Principles and Practice of Multi-agent Systems. Nagoya: Springer, 2020. 343–351.

(校对责编: 孙君艳)