对手类型未知情况下的两人零和马尔科夫博弈决策

王成意, 朱　进†, 赵云波

(中国科学技术大学 信息科学技术学院, 安徽 合肥 230026)

摘要: 本文研究一类典型的非完全信息博弈问题—对手类型未知的两人零和马尔科夫博弈, 其中对手类型多样且每次博弈开始前无法得知对手类型. 文中提出了一种基于模型的多智能体强化学习算法—对手辨识的极大极小Q学习(DOMQ). 该算法首先建立对手相关环境的经验模型, 再使用经验模型学习纳什均衡策略, 己方智能体在实际博弈中根据经验模型判断对手类型, 从而使用相应的纳什均衡策略, 以保证收益下限. 本文所提的DOMQ算法只需要在采样阶段的每轮博弈结束后得知对手的类型, 除此之外无需知道任何环境的信息. 仿真实验验证了所提算法的有效性.

关键词: 两人零和马尔科夫博弈; 非完全信息; 极大极小Q学习; 纳什均衡; 多智能体强化学习

引用格式: 王成意, 朱进, 赵云波. 对手类型未知情况下的两人零和马尔科夫博弈决策. 控制理论与应用, 2024, 41(11): 2131 – 2138

DOI: 10.7641/CTA.2023.20630

Decision making for two-player zero-sum Markov games with indistinguishable opponents

WANG Cheng-yi, ZHU Jin†, ZHAO Yun-bo

(School of Information Science and Technology, University of Science and Technology of China, Hefei Anhui 230026, China)

Abstract: This paper investigates a typical class of incomplete information games – two-player zero-sum Markov games with indistinguishable opponents, where the opponent types are diverse and cannot be known at the beginning of the game. We propose a model-based multi-agent reinforcement learning algorithm – distinguishing opponent minimax Q-learning (DOMQ). The algorithm firstly builds an empirical model of the opponent-related environment; secondly uses the empirical model to learn a Nash equilibrium strategy, and then uses the corresponding Nash equilibrium strategy to guarantee the lower bound of the return in actual game. All the necessary information needed for the proposed DOMQ algorithm is the opponent type at the end of each episode in the sampling period rather than the other information about the environment. The simulation results verify the effectiveness of the proposed algorithm.

Key words: two-player zero-sum Markov game; incomplete information; minimax Q-learning; Nash equilibrium; multi-agent reinforcement learning

Citation: WANG Chengyi, ZHU Jin, ZHAO Yunbo. Decision making for two-player zero-sum Markov games with indistinguishable opponents. *Control Theory & Applications*, 2024, 41(11): 2131 – 2138

1 引言

马尔科夫决策过程(Markov decision processes, MDPs)[1] 是描述单智能体与环境交互的常用模型. 现实中的大量场景均可通过MDPs进行建模并利用强化学习来求取相应的最优决策[2–3]. 然而, 当环境中存在多个智能体时[4–5], 各个智能体的动作会对其他智能体产生影响, 同时其他智能体的动作也会改变该智能体获得的奖励, 从而使得基于MDPs的强化学习无法处理该类场景中的问题[6]. 在这种情况下, 原先针对单智能体的MDPs模型被推广至多智能体环境, 称为马尔科夫博弈 (Markov games, MGs)[7], 或随机博弈 (stochastic games)[8].

作为MGs的一种特定形式, 两人零和马尔科夫博弈(two-player zero-sum Markov games, TZMGs)广泛分布于现实生活中, 它包含两个智能体并且每个智能体获得的奖励互为相反数[9–10]. 博弈根据参与者的执

收稿日期: 2022−07−15; 录用日期: 2023−09−11.

†通信作者. E-mail: jinzhu@ustc.edu.cn; Tel.: +86 18956012051.

本文责任编委: 王龙.

国家重点研发计划项目(2018AAA0100802), 安徽省自然科学基金项目(2008085MF198)资助.

Supported by the National Key Research and Development Program (2018AAA0100802) and the National Science Foundation of Anhui Province (2008085MF198).

行动作是否有顺序可以分为两类: 一类是基于次序的博弈(turn-based games), 这种博弈的参与者执行动作是有先后顺序的, 可以用一棵博弈树来描述, 树中的每个节点表示博弈进行中的每一个可能的状态, 博弈从唯一的初始节点开始进行, 通过由参与者决定的路径到达终端节点, 此时博弈结束, 参与者得到相应的收益, 如围棋[11]、两人德州扑克[12–13]等; 与基于次序的博弈相对应的是同步博弈(simultaneous games), 博弈的参与者在同一时间行动, 因此无法得知其他参与者将会采取什么动作, 这也使得做出决策更加困难, 如无人机对抗[4,14]、双人对抗游戏[15]等. 作为经典的基于次序的博弈, 两人德州扑克问题近年来被广泛研究, 反事实后悔值最小化(counterfactual regret minimization, CFR)算法[16]是解决德扑问题最有效的手段, 文献[12–13]均使用了该算法, 并达到了超越顶尖人类玩家的性能. 而由于CFR需要博弈可以展开为博弈树的限制, 其并不适用于同步博弈. 文献[17]提出了一种结合神经网络和极大极小Q学习的在线学习算法, 求解同步TZMGs的纳什均衡策略. 文献[18]提出了一种学习自动机(learning automata, LA)解决方案, 在使用纯策略时可能没有纳什均衡的情况下, 它也能收敛到混合策略纳什均衡. 本文主要关注同步的TZMGs.

针对智能体是否得知全部信息这一特点, 博弈可以分为完全信息博弈和非完全信息博弈[19]. 相较于完全信息博弈, 非完全信息博弈中存在部分信息(如环境或对手信息)未知的情况, 这导致玩家难以预估可以得到的奖励, 因此更具挑战性. 在非完全信息的诸多体现形式中, 对手类型多样化且未知是重要的一种. 例如在网络游戏中, 游戏AI智能体面对的对手可能有竞技型、娱乐型等多种类型, 而AI智能体的收益是对手的满意度, 即使是同样的对局, 不同类型的玩家满意度可能不同, 因此使AI智能体的收益不同; 或者在无人机对抗中, 敌方无人机可能有多种型号, 不同型号的无人机使用相同的动作(如开火)可能会产生不同的效果, 从而导致我方无人机获得不同的收益. 在以上情况中, 己方智能体无法通过对方的动作或者外观来判别对手的真实类型. 这些博弈场景的共通点是博弈开始时己方智能体无法确定对手的类型及环境的参数(如转移概率和奖励函数)等, 因此己方智能体无法生成有效决策以获取最大奖励. 为了简化问题, 本文主要关注上述博弈场景的两人零和形式.

本文将该类问题建模为一个对手类型未知的两人零和马尔科夫博弈模型(opponent-indistinguishable TZMGs, OI-TZMGs), 提出了一种基于模型的多智能体的强化学习算法—对手辨识的极大极小Q学习(distinguishing opponent mini-max Q-learning, DOMQ). 该方法从收集到的数据中构建对手相关环境模型并使用相应的环境模型学习纳什均衡策略, 然后在实际博弈中通过博弈历史分辨对手类型, 并据此使用相应的策略以保证收益下限. 最后, 本文根据OI-TZMGs的性质搭建了两个仿真环境, 并在这两个环境中分别使用DOMQ、信息完全已知的极大极小Q学习和传统极大极小Q学习进行验证. 实验结果表明, 针对OI-TZMGs问题, DOMQ算法性能接近信息完全已知的极大极小Q学习算法, 并且明显优于传统极大极小Q学习.

2 问题描述

2.1 对手类型未知的两人零和马尔科夫博弈

一个包含M种类型对手的步长有限的两人零和马尔科夫博弈模型可以用一个八元组$\{\mathcal{S},\mathcal{A},\mathcal{B},\mathcal{P},\mathcal{R},\gamma,\{\beta_m\}_{m=1}^M,\{\omega_m\}_{m=1}^M\}$表示. 博弈开始时, 初始状态为$s_0\in\mathcal{S}$, 其中$\mathcal{S}$为该博弈的状态空间. 博弈的对手$\{\beta_m\}_{m=1}^M$以概率$\{\omega_m\}_{m=1}^M$出现, 即该次博弈的对手为$\beta_m$的概率为$\omega_m$, 在博弈结束前, 己方玩家无法得知对手类型. 己方玩家和敌方玩家在$t(t=0,1,2,\cdots)$时刻分别从自己的动作空间$\mathcal{A}$和$\mathcal{B}$中执行一个动作a_t,b_t, 环境状态根据转移函数$\mathcal{P}$转移到下一状态s_{t+1}, 其概率为$\mathcal{P}(s_{t+1}|s_t,a_t,b_t)$, 同时己方玩家根据奖励函数$\mathcal{R}$获得$r_{t+1}$奖励, 敌方玩家的奖励则为$-r_{t+1}$. 奖励函数$\mathcal{R}$是一个与对手类型相关的映射: $(s_t,a_t,b_t,s_{t+1},\beta)\to\Delta(R)$, 其中$R$是己方奖励的奖励空间, $\Delta(R)$是R上的一个概率分布, 己方玩家获得r_{t+1}奖励的概率为$\mathcal{R}(r_{t+1}|s_t,a_t,b_t,s_{t+1},\beta)$. 己方玩家的目标是最大化自己期望的累计奖励(也称为回报)

$$V(s_0)=\mathrm{E}[\sum_{t=0}^{\infty}\gamma^t r_{t+1}],\tag{1}$$

其中$\gamma\in[0,1)$为折扣因子, 由于是零和博弈, 敌方玩家的目标则是最小化上述期望回报. 本文将上述博弈称为对手类型未知的两人零和马尔科夫博弈.

注 1 己方玩家无论面对哪种类型的对手, 奖励空间都是一样的, 但获得奖励的概率分布可能不一样. 即对于不同类型的对手β_i,β_j, 可能有

$$\mathcal{R}(s_t,a_t,b_t,s_{t+1},\beta_i)\neq\mathcal{R}(s_t,a_t,b_t,s_{t+1},\beta_j),i\neq j,\tag{2}$$

正是由于OI-TZMGs的这个性质, 才导致博弈的对手难以区分. 特别地, 当该概率分布发生退化时(即奖励函数为与对手类型相关的实值函数), 则可以通过查看奖励来判断对手的类型, 这种情况下对手不难区分.

2.2 纳什均衡

在单智能体环境中, 智能体的目标是获得最大的累计奖励, 而累计奖励取决于智能体在每个状态下选择的动作, 这样的动作选择方式被称为策略. 策略一般用π表示, 策略π在状态s使用动作a的概率记为$\pi(a|s)$. 使得智能体获得最大累计奖励的策略称为最优策略.

不同于单智能体环境, 在一般的MGs中, 没有玩家可以获得最优策略, 原因在于玩家的奖励不仅与自己的动作有关, 也和其他玩家的动作有关, 而其他玩家的策略并不是固定的, 因此策略无法收敛到最优. 博弈论中存在两个经典概念——最佳响应(best response, BR)[20]和纳什均衡(Nash equilibrium, NE)[21], 本文使用它们来说明在MGs中期望达成的目标.

定义 1 (最佳响应)　在一个MG中, 给定其他玩家的策略π^{-i}, 其中$-i$表示除玩家i之外的所有玩家, 若玩家i的策略π_b^i可以使自己获得最大的回报, 即满足以下等式:

$$V(s_0;\pi_b^i,\pi^{-i})=\max_{\pi^i}V(s_0;\pi^i,\pi^{-i}),\tag{3}$$

则称π_b^i是对π^{-i}的最佳响应, 记作$\pi_b^i=BR(\pi^{-i})$. 其中$V(s_0;\pi^i,\pi^{-i})$ 表示玩家i使用策略π^i, 其他玩家使用策略π^{-i}时玩家i所获得的回报.

定义 2 (纳什均衡)　在一个 MG 中, 若每个玩家的策略都是对其他玩家策略组合的最佳响应, 即对于任意玩家 i, 其策略为 π_*^i, 其他玩家的策略为 π_*^{-i}, 都有

$$V(s_0;\pi_*^i,\pi_*^{-i})=\max_{\pi^i}V(s_0;\pi^i,\pi_*^{-i}),\tag{4}$$

则称所有玩家的策略组合$(\pi_*^1,\pi_*^2,\cdots,\pi_*^n)$为纳什均衡, 记作$\sigma_*$.

在TZMGs中, 纳什均衡总是存在的, 设双方玩家的纳什均衡策略组为(μ_*,ν_*), 则己方玩家的回报可表示为

$$V(s_0;\mu_*,\nu_*)=\max_{\mu}\min_{\nu}V(s_0;\mu,\nu)=\min_{\nu}\max_{\mu}V(s_0;\mu,\nu).\tag{5}$$

在纳什均衡中, 博弈中的任意玩家无法通改变自己策略获得更高回报. 特别地, 在TZMGs中, 当己方玩家使用纳什均衡策略时, 无论对手使用何种策略, 其回报不会优于使用纳什均衡策略. 这表明即使对手可以一直学习, 不断优化自己的策略, 己方智能体也可以通过使用纳什均衡策略来保证自己回报的下限. 由于本文对对手的策略没有进行任何假设, 因此学习纳什均衡策略保证回报下限具有很强的现实意义.

3　对手辨识的极大极小Q学习

3.1　对手相关环境建模

由于己方智能体需要学习到针对每个对手的纳什均衡策略, 而每轮博弈开始前己方智能体无法判断对手的类型, 于是本文的算法希望建立对手相关的环境模型, 再使用该模型学习针对该对手的纳什均衡策略. 模型一般包含两个部分: 转移概率和奖励函数. 若OI–TZMGs中包含M种不同类型的对手, 则算法需要建立M个转移概率和奖励函数模型. 在学习纳什均衡策略时, 使用相应的模型来进行训练, 在实战中, 通过模型判断对手的类型, 使用相应的纳什均衡策略应对. 在本文中, “对手相关模型”指的便是该对手对应的转移概率和奖励函数模型.

本文对环境建模做出如下假设:

假设 1　环境的生成模型(即真实模型)$\mathcal{G}$存在;

假设 2　每轮博弈结束后得知对手类型.

拥有以上假设后, 对于任意一个状态动作元组(s,a,b), 将其输入到生成模型$\mathcal{G}$中, 都可以得到下一步的奖励和状态, 并且在每轮博弈结束后都得知对手的类型. 因此算法可以使用生成模型对每个状态动作元组(s,a,b)进行N次采样, 然后通过采样得到的数据构建真实模型$\mathcal{G}$. 本文将通过数据构建的模型称为经验模型, 记为$\hat{\mathcal{G}}$. 分别令$\hat{P}_m,\hat{R}_m$为$\hat{\mathcal{G}}$的第m种类型对手的转移概率和奖励函数, 记为$\hat{\mathcal{G}}_m$, 对于状态和动作都有限的环境, 可以通过

$$\begin{aligned}\hat{P}_m(s'|s,a,b)&=\frac{\mathrm{count}(s',s,a,b)}{N},\\ \hat{R}_m(r|s,a,b,s')&=\frac{\mathrm{count}(r,s,a,b,s')}{\mathrm{count}(s',s,a,b)},\end{aligned}\tag{6}$$

计算得到$\hat{P}_m,\hat{R}_m$. 其中$\mathrm{count}(s',s,a,b)$表示在状态s下使用联合动作(a,b), 状态转移到s'的次数, $\mathrm{count}(r,s,a,b,s')$表示在状态$s$下使用联合动作$(a,b)$, 状态转移至$s'$时获得奖励为$r$的次数, N为在状态动作元组(s,a,b)采样的次数.

当环境的状态和动作都有限时, 本文使用采样策略

$$\pi_s(s_t)=\arg\min_{a_t}\mathrm{count}(s_t,a_t,b_t),\tag{7}$$

对生成模型$\mathcal{G}$进行采样, 采样得到的奖励r_{t+1}和下一个状态s_{t+1}表示为

$$(r_{t+1},s_{t+1})=\mathcal{G}(s_t,a_t,b_t),\tag{8}$$

其中$a_t=\pi_s(s_t)$, 直到所有状态动作元组(s,a,b)的采样数都大于等于某个设定值.

对于状态或动作无限的环境, 文献[22]提出了一种基于轨迹采样的概率估计方案, 将深度网络应用于模型估计. 因此, 可以使用该方法得到第m个对手相关的经验模型$\hat{\mathcal{G}}_{\theta_m}$. 可以构建概率神经网络, 其输出神经元为简单概率分布函数的参数, 如高斯分布的期望和方差, 或均匀分布的上下界等, 以描述环境或数据带来的不确定性. 本文使用负对数预测概率作为损失函数, 即

$$\mathrm{loss_P}(\boldsymbol{\theta})=-\sum_{n=1}^{N}\log\widetilde{f}_{\boldsymbol{\theta}}\left(\boldsymbol{s}_{n+1}\mid\boldsymbol{s}_n,\boldsymbol{a}_n\right).\tag{9}$$

以最为常见的高斯分布为例, 以θ为参数的神经网络可表示为$f=\Pr(\boldsymbol{s}_{t+1}\mid\boldsymbol{s}_t,\boldsymbol{a}_t)=\mathcal{N}(\boldsymbol{\mu}_\theta(\boldsymbol{s}_t,\boldsymbol{a}_t),$

$\boldsymbol{\Sigma}_\theta(\boldsymbol{s}_t,\boldsymbol{a}_t)$), 此时, 损失函数为

$$\begin{aligned}\text{loss}_{\text{Gauss}}(\boldsymbol{\theta})=&\sum_{n=1}^{N}[\mu_{\boldsymbol{\theta}}(\boldsymbol{s}_n,\boldsymbol{a}_n)-\boldsymbol{s}_{n+1}]^{\text{T}}\times\\&\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{s}_n,\boldsymbol{a}_n)[\mu_{\boldsymbol{\theta}}(\boldsymbol{s}_n,\boldsymbol{a}_n)-\boldsymbol{s}_{n+1}]+\\&\log\det\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{s}_n,\boldsymbol{a}_n).\end{aligned}\tag{10}$$

这样的神经网络可以预测环境的不确定性, 在对环境噪声不确定的情况下, 高斯分布通常来说是比较稳妥的选择. 此后再通过梯度下降算法优化损失函数, 以训练网络模型, 直到网络可以较为准确地预测下一个状态. 此时, 第m个对手相关的环境经验模型可以由以θ为参数的神经网络表示, 即

$$(r_{t+1},s_{t+1})=\hat{\mathcal{G}}_{\theta_m}(s_t,a_t,b_t).\tag{11}$$

3.2 极大极小Q学习

在获得了每个对手相关环境模型后, 己方智能体可以使用这些模型, 学到应对每个对手的纳什均衡策略. 本小节通过使用极大极小Q学习的方法来学习应对每个对手的纳什均衡策略.

3.2.1 求解纳什均衡

定理 1 TZMGs的纳什均衡Q^*值是唯一的.

证 首先, 根据MDPs中状态价值函数V^*的贝尔曼最优方程

$$V^*(s)=\max_{\pi(s)\in\Pi}\sum_{a\in\mathcal{A}}\pi(a\mid s)Q^*(s,a),\tag{12}$$

以及纳什均衡的定义, 可以得到TZMGs中状态价值函数V^*的表达式

$$V^*(s)=\max_{\pi(s)\in\Pi}\min_{b\in\mathcal{B}}\sum_{a\in\mathcal{A}}\pi(a\mid s)Q^*(s,a,b),\tag{13}$$

同理可得状态动作价值函数Q^*在TZMGs中的表达式

$$\begin{aligned}Q^*(s,a,b)=&\,\mathcal{R}(s,a,b)+\\&\gamma\sum_{s'\in\mathcal{S}}\mathcal{P}\left(s'\mid s,a,b\right)V^*\left(s'\right).\end{aligned}\tag{14}$$

将式(14)代入式(13), 可得

$$\begin{aligned}V^*(s)=&\max_{\pi(s)\in\Pi}\min_{b\in\mathcal{B}}\sum_{a\in\mathcal{A}}\pi(a\mid s)[\mathcal{R}(s,a,b)+\\&\gamma\sum_{s'\in\mathcal{S}}\mathcal{P}\left(s'\mid s,a,b\right)V^*\left(s'\right)],\end{aligned}\tag{15}$$

上式被称为V^*值的贝尔曼极大极小方程.

同理, 将式(13)代入式(14), 可以得到Q^*值的贝尔曼极大极小方程

$$\begin{aligned}Q^*(s,a,b)=&\,\mathcal{R}(s,a,b)+\gamma\sum_{s'\in\mathcal{S}}\mathcal{P}\left(s'\mid s,a,b\right)\times\\&\max_{\pi(s)\in\Pi}\min_{b'\in\mathcal{B}}\sum_{a'\in\mathcal{A}}\pi\left(a'\mid s'\right)Q^*\left(s',a',b'\right),\end{aligned}\tag{16}$$

根据沙普利(Shapley)理论[8], 等式(16)存在唯一解Q^*. 证毕.

注 2 在式(15)–(16)中, $\pi(s)$表示给定状态s下动作的分布, Π表示在己方动作集$\mathcal{A}$上所有可能分布的集合, $\pi(a|s)$表示在策略π下状态s时使用动作a的概率. 不同于MDPs中的贝尔曼最优方程, 贝尔曼极大极小方程中最大化算子并不只取单一的动作, 而是取动作集上的一个分布, 这是因为在TZMGs中不一定存在纯策略纳什均衡, 但一定存在混合策略纳什均衡[23]. 在这里, 纯策略属于混合策略的一种特殊情况. 而最小化算子则是取单一的动作, 因为当最大化算子确定了一个策略后, 对于对手来说, 这就退化为一个单智能体问题, 因此一定存在纯策略使得价值函数最小, 为简化计算, 即可取纯策略(即单一动作).

定理 2 若已知纳什均衡Q^*值, 则可以通过求解线性规划问题

$$\begin{cases}\max & c\\ \text{s.t.} & \sum_a p(a)Q^*(s,a,b)\geqslant c,\ \forall b\in\mathcal{B},\\ & \sum_a p(a)=1,\\ & p(a)\geqslant 0,\ \forall a\in\mathcal{A},\end{cases}\tag{17}$$

得到纳什均衡V^*值及纳什均衡策略.

证 当已知Q^*时, 根据式(13), 可以通过

$$\pi^*(s)=\arg\max_{\pi}\min_{b}\sum_{a}\pi(a\,|\,s)Q^*(s,a,b)\tag{18}$$

计算纳什均衡策略, 则式(17)中的c即是纳什均衡的V^*值, $p(a)$则是相应的智能体A在状态s下的纳什均衡策略, 即$p(a)=\pi(a|s)$. 证毕.

3.2.2 求解纳什均衡Q^*值

根据定理2, 一旦求出纳什均衡的Q^*值, 就可以通过求解式(17)得到纳什均衡策略. 所以问题转换为了如何求解Q^*. 自然的想法是通过式(16)求解, 然而, 该式的最大化操作中的π是连续的, 使得难以直接使用该式求解Q^*. 单智能体强化学习中的Q学习是通过迭代的方式计算Q值, 可以使用相同的思想, 用Q学习来近似计算极大极小方程中的Q^*, 这称为极大极小Q学习[7], 根据定理1, 所求的Q^*是唯一的.

然而, Q学习是一种表格形式的学习方法, 只适用于状态和动作有限的情况. 单智能体强化学习中为了解决这一问题, 通过引入神经网络来近似计算任意状态下的Q值, 将Q学习推广到状态和动作无限的情况, 这种技术称为深度Q网络(deep Q network, DQN)[2]. 同理, DQN也可以和极大极小Q学习相结合, 以解决状态和动作无限的情况. 文献[17]提出的方法可以应用于该场景.

3.3 对手辨识

在得到面对各个对手的纳什均衡策略后, 己方智能体需要在每轮博弈开始后尽快分辨出本次博弈面对的对手, 以使用相应的策略. 在第3.1节中已经得到了对手相关的环境模型, 因此, 通过分析与对手交互的历史, 即可判断是在和哪个类型的对手进行博弈.

具体来说, 对于状态和动作有限的博弈, 算法记录与对手交互的历史h, 所有历史集合记为$\mathcal{H}$, $h \in \mathcal{H}$. 对于一个长度为n的历史$h=\{(s_0,a_0,b_0,r_1,s_1),(s_1,a_1,b_1,r_2,s_2),\cdots,(s_{n-1},a_{n-1},b_{n-1},r_n,s_n)\}$, 已知经验模型为$\hat{\mathcal{G}}$的时候, 可以计算出对手是第$m$种类型的条件下产生$h$的概率为

$$p(h|m)=\prod_{i=0}^{n-1}[\hat{P}_m(s_{i+1}|s_i,a_i,b_i)\times \hat{R}_m(r_{i+1}|s_i,a_i,b_i,s_{i+1})], \tag{19}$$

那么, 可以通过

$$j=\arg\max_m p(h|m), \tag{20}$$

确定对手的类型β_j. 对于状态或动作无限的情况, 式(19)中的$\hat{P}_m(s_{i+1}|s_i,a_i,b_i)$和$\hat{R}_m(r_{i+1}|s_i,a_i,b_i)$可由估计模型$\hat{\mathcal{G}}_{\theta_m}$得出. 对于状态和动作有限的环境, DOMQ的整体算法见表1.

表 1 DOMQ: 分辨对手的极大极小Q学习

Table 1 Distinguishing opponent minimax Q-learning

输入: 生成模型$\mathcal{G}$, 对手数量M.

输出: 智能体A使用纳什均衡策略的回报.

Initialize: 智能体A的策略$\pi_{\rm A}^1,\pi_{\rm A}^2,\cdots,\pi_{\rm A}^M$, 经验模型$\hat{\mathcal{G}}$, 采样策略$\pi_s$;
while $n<N$ **do**
  初始化生成模型$\mathcal{G}$, 对手类型为m;
  初始化对手的策略$\pi_{\rm B}^m(s)$;
  while 博弈未结束 **do**
    $a_t=\pi_s(s_t), b=\pi_{\rm B}^m(s)$;
    $(r_{t+1},s_{t+1})=\mathcal{G}(s_t,a_t,b_t)$;
    将$(s_t,a_t,b_t,r_{t+1},s_{t+1})$存入$\mathcal{D}_{\rm env}$;
  end while
  获得对手类型m, $\mathcal{D}_{\rm env}^m=\mathcal{D}_{\rm env}^m+\mathcal{D}_{\rm env}$, $\mathcal{D}_{\rm env}=\emptyset$;
  $n=\min \text{count}(s_t,a_t,b_t)$;
end while
for $m=1,2,\cdots,M$ **do**
  用式(6)计算第m个对手的经验模型$\hat{\mathcal{G}}_m$;
end for
获得全部经验模型$\hat{\mathcal{G}}$;
for $m=1,2,\cdots,M$ **do**
  使用极大极小Q学习和经验模型$\hat{\mathcal{G}}$计算应对对手m的纳什均衡策略$\pi_{\rm A}^m$;
end for
初始化历史h, 对手类型m, 对手策略$\pi_{\rm B}^m$;
while 博弈未结束 **do**
  使用$\hat{\mathcal{G}}$, h和式(19)–(20)分辨对手, 得到对手类型j;
  $a=\pi_{\rm A}^j(s), b=\pi_{\rm B}^m(s)$, 用$(a,b)$进行博弈;
end while

4 实验

本文根据OI-TZMGs的性质搭建了两个仿真环境, 分别是状态空间有限的网格世界和状态空间无限的圆环游戏, 以验证本文算法的有效性.

4.1 网格世界

在本文的网格世界仿真环境中, 总共有9个状态, 如图1(a)所示, 在这个环境中有两种不同类型的对手, 即对手1和对手2, 每局游戏开始时, 会均匀随机地选择一个对手. 面对对手1时, 己方智能体的奖励函数如图1(b)所示, 图中的数字表示己方智能体获得1奖励的概率, 反之则获得−1奖励, 例如在状态1, 己方智能体面对对手1, 有0.3的概率获得1奖励, 0.7的概率获得−1奖励. 同理, 图1(c)为己方智能体面对对手2的奖励函数. 在这个网格世界中, 己方智能体和对手智能体的动作空间均为{上, 停, 下}, 环境根据双方的动作按一定规则转移到下一状态, 例如双方玩家都选择“上”时, 状态以0.6 的概率向上移动, 以0.4的概率随机移动, 同时双方智能体获得自己的奖励.

值得一提的是, 环境的状态转移是可以越过边界的, 例如在状态8向上移动, 则会移动到状态2. 具体转移概率见表2.

首先, 本文的算法使用采样策略收集游戏的数据, 并根据对手1和对手2的游戏对局使用式(6) 分别建立它们的相关环境模型; 然后, 算法使用对手1和对手2的环境模型, 用极大极小Q学习学习了面对对手1和对

7	8	9
4	5	6
1	2	3

(a) 各网格的状态

0.7	0.7	0.7
0.5	0.5	0.5
0.3	0.3	0.3

(b) 己方智能体面对对手1的奖励函数

0.3	0.3	0.3
0.5	0.5	0.5
0.7	0.7	0.3

(c) 己方智能体面对对手2的奖励函数

图 1 网格世界

Fig. 1 Gridworld

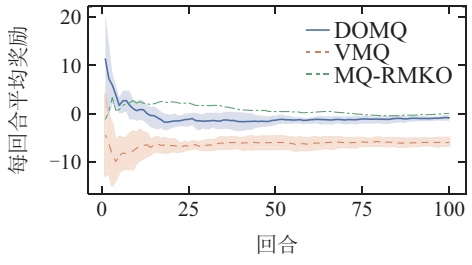
手2的纳什均衡策略, 最后, 本文测试了DOMQ算法在实际环境中的效果, 并使用了其他两种算法——经典极大极小Q学习 (vanilla minimax Q-learning, VMQ)和已知真实环境和对手的极大极小Q学习(minimax Q-learning with real model and known opponent, MQ-RMKO)进行对比, 如图2所示. 由于环境的随机性较大, 每次实验都重复进行5次, 图中的实线表示5次实验的平均值, 阴影部分表示5次实验的标准差.

表 2 网格世界转移概率

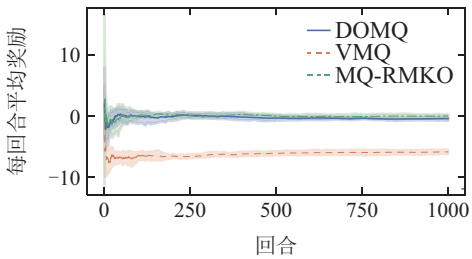
Table 2 Transactions of gridworld

联合动作	向上	向下	停止	随机移动
(上, 上)	0.6	0	0	0.4
(上, 停)	0.4	0	0	0.6
(上, 下)	0	0	0.4	0.6
(停, 上)	0.4	0	0	0.6
(停, 停)	0	0	0.4	0.6
(停, 下)	0	0.4	0	0.6
(下, 上)	0	0	0.4	0.6
(下, 停)	0	0.4	0	0.6
(下, 下)	0	0.6	0	0.4

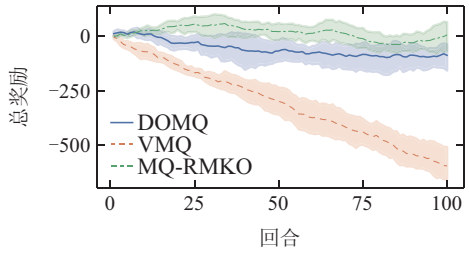
值得说明的是, MQ-RMKO方法是完全已知游戏的所有信息, 包括每局游戏开始时的对手类型, 所以其策略组可快速地收敛到纳什均衡. 从理论上分析, 本文的方法DOMQ的性能肯定不如MQ-RMKO方法, 因为本文的方法在仅知道对手有几种类型, 而不知道任何其他游戏信息(如转移函数, 对手相关的奖励模型, 每局游戏开始时对手的类型). 然而即便如此, DOMQ展现了优秀的性能. 在图2(a)中, DOMQ算法100轮次的平均回报十分接近MQ-RMKO方法, 而显著优于VMQ.



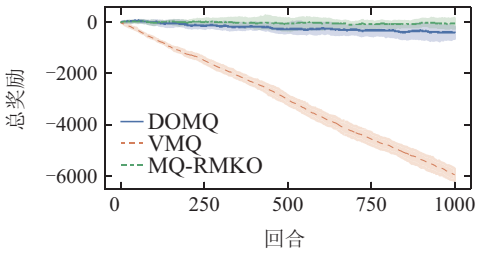
(a) 100轮游戏的平均回报



(b) 1000轮游戏的平均回报



(c) 100轮游戏的累计回报



(d) 1000轮游戏的累计回报

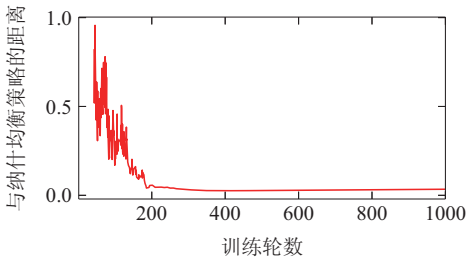
图 2 网格世界智能体A的回报

Fig. 2 Reward of agent A in gridworld

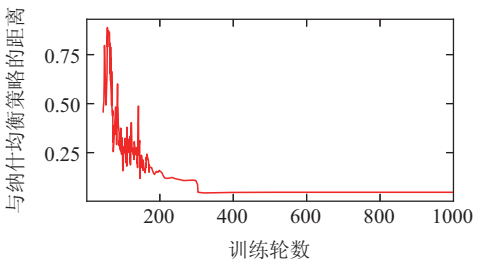
在图2(b)中也展现了类似的性能, 己方智能体的每轮平均回报只有不到–0.5, 而VMQ则是达到了–6左右. 在累计回报上(图2(c)–(d)), DQMQ算法有着相似的表现, 在不知道游戏的诸多信息的情况下达到了接近最优的性能.

此外, 本文还在网格世界中验证了基于模型的情况下, 本文的算法能否收敛到纳什均衡策略, 本文在得到了对手相关环境模型后, 使用相应模型学习纳什均衡策略, 如图3所示, 横轴表示学习的轮次, 纵轴表示当前策略与纳什均衡策略之间的距离, 图3中与纳什均衡策略的距离定义为

$$d = \frac{1}{|\mathcal{S}|}\sum_{i=1}^{|\mathcal{S}|}\|\pi(s_i) - \pi_*(s_i)\|_2. \tag{21}$$



(a) 策略与纳什均衡策略的距离(对手类型为1)



(b) 策略与纳什均衡策略的距离(对手类型为2)

图 3 智能体A策略与纳什均衡策略的距离

Fig. 3 Distance to Nash policy of agent A

由图3可见, 本文的算法在面对两种不同的对手时, 均只需要大约250轮离线训练, 策略即可收敛至近似纳什均衡策略.

4.2　圆环游戏

由于网格世界的状态是离散的, 本文另外搭建了一个状态是连续的仿真环境——圆环游戏, 在这个环境中也是存在两种不同类型的对手, 每种对手出现的概率相等, 图4展示了面对不同类型对手的奖励函数, 其中左侧的圆环表示面对对手1时的奖励函数, 右侧的圆环表示面对对手2时的奖励函数, 圆环的指针指向的地方即为己方获得奖励1的概率, 例如面对对手1时, 圆环的指针指向0.6, 则此时己方获得奖励1的概率为0.6, 获得奖励−1的概率为0.4.

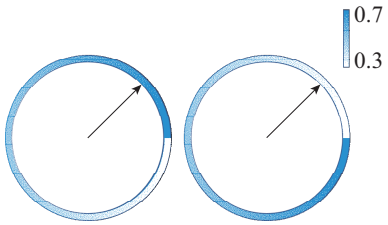


图 4　圆环游戏智能体A的奖励函数

Fig. 4　Reward function of agent A in annulus game

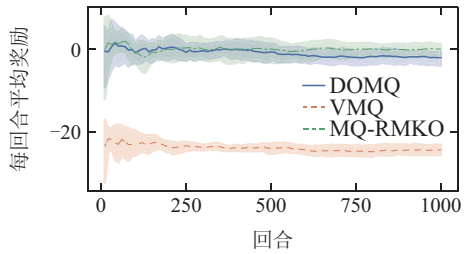
指针与水平的夹角即为环境的状态, 因此环境的状态空间为$[0, 360)$, 是一个连续的空间. 己方玩家和对手的动作空间一样, 均为{顺, 停, 逆}, 顺代表顺时针拨动指针, 逆代表逆时针拨动指针, 停则表示不动. 环境的状态按以下公式转移:

$$s' = (s + r(30) \times c(0) - r(30) \times c(2)) \bmod 360.$$

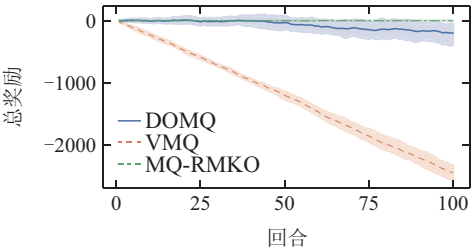
其中: $r(x)$表示均匀随机生成一个$[0, x]$之间的数, $c(0)$表示动作“顺”出现的次数, $c(2)$表示动作“逆”出现的次数, s为当前状态, s'为双方执行完动作后的下一状态.

在圆环游戏中, 和网格世界的实验步骤类似, 在为对手相关环境模型建模时, 本文搭建了两个概率神经网络对环境的转移函数和奖励函数分别进行建模, 转移函数网络为$3 \times 128 \times 32 \times 4$的网络, 其输入是当前状态——动作元组$(s_t, a_t, b_t)$, 输出则是下一个状态的概率分布类别及参数.

本文由收集的数据分析, 在双方玩家联合动作属于{(顺, 停), (逆, 停), (停, 顺), (停, 逆)}时, 状态的增量大概率服从均匀分布, 当联合动作属于{(顺, 顺), (逆, 逆), (逆, 顺), (顺, 逆)}时, 状态的增量大概率服从三角分布, 因此神经网络输出4个参数: $(i, \text{upper}, \text{lower}, \text{mid})$, 其中$i$表示状态增量服从的分布类型, 0代表无分布, 1代表均匀分布, –1代表三角分布, upper表示分布的上界, lower表示分布的下界, mid表示三角分布的中间值. 神经网络输出这4个参数, 通过简单的运算即可预测下一个状态及其概率密度. 奖励函数网络则是$4 \times 128 \times 32 \times 1$的网络, 其输入为四元组$(s_t, a_t, b_t, s_{t+1})$, 输出为智能体A获得奖励1的概率. 在训练好这两个神经网络后, 便得到了对手相关环境模型. 然后使用对手1和对手2的环境模型, 将状态离散化, 离散后的状态$s^* = \lfloor s \rfloor$, 因此无限状态空间被离散化为有限状态空间, 仍可以用极大极小Q学习来学习纳什均衡策略; 最后, 在真实环境中测试了本文算法在无限状态环境中的有效性, 实验结果见图5. 和网格世界的实验结果类似, 圆环游戏中DOMQ的性能也十分接近MQ-RMKO, 这得益于概率神经网络对环境的精准建模.



(a) 100轮游戏的平均回报



(b) 100轮游戏的累计回报

图 5　圆环游戏智能体A的回报

Fig. 5　Reward of agent A in annulus game

4.3　结果分析

从实验结果上看, 本文提出的DOMQ算法的性能显著优于VMQ. DOMQ中己方智能体的策略都是在对手相关的环境中学习到的, 对每种类型的对手都有一定的针对性, 而VMQ 则是不知道环境中有若干个类型的对手, 或者说即使知道也无法区分, 自然难以学习到应对每个对手的纳什均衡策略. VMQ只能通过与环境交互, 得到一个平均状态下的策略, 这在对手对环境影响很小的时候, 也许有不错的性能, 但一旦对手对环境的影响较大, 那么性能会急速下降. 此外, DOMQ 算法还实现了接近MQ-RMKO算法的性能, 这是十分优秀的结果, MQ-RMKO已知环境所有信息和每局游戏对手的类型, 这意味着在MQ-RMKO方法中, 博弈变成了完全信息的博弈. 非完全信息博弈比完全信息博弈困难得多, 而DOMQ仍能达到接近MQ-RMKO的性能, 也证明了本文提出的方法的有效性. 而DOMQ与MQ-RMKO之间的差距, 主要由两个方面产生: 1) 在游戏历史较少的时候,辨别对手的类型可能会出错; 2) 在建模阶段收集到的数据有限,

建立的模型存在误差. 这两个问题在本文的OI-TZMGs环境中是无法避免的, 因此和MQ-RMKO方法有一定差距在所难免.

5 结论

本文针对非完全信息博弈中的一个常见问题—存在多种对手类型且难以区分, 提出了一种更为宽泛OI-TZMGs模型, 并给出了对手辨识的极大极小Q学习(DOMQ). 该算法首先建立对手相关的环境模型, 再使用相应模型训练己方智能体以获得与相应对手的纳什均衡策略, 最后在实际对战中判断对手类型, 使用相应的策略以达到均衡. 在本文搭建的仿真环境中, DOMQ算法达到了接近信息完全已知的极大极小Q学习算法的性能, 并且性能显著优于传统的极大极小Q学习.

参考文献:

[1] HOWARD R A. *Dynamic Programming and Markov Processes*. Hoboken, NJ, USA: Wiley, 1960: 17 – 21.

[2] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning. *Nature*, 2015, 518(7540): 529 – 533.

[3] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning. *International Conference on Machine Learning*, New York, USA: PMLR, 2016, 48: 1928 – 1937.

[4] KONG Weiren, ZHOU Deyun, ZHAO Yiyang, et al. Maneuvering strategy generation algorithm for multi-UAV in close-range air combat based on deep reinforcement learning and self-play. *Control Theory & Applications*, 2022, 39(2): 352 – 362.
(孔维仁, 周德云, 赵艺阳, 等. 基于深度强化学习与自学习的多无人机近距空战机动策略生成算法. 控制理论与应用, 2022, 39(2): 352 – 362.)

[5] BAO Tao, LI Haofei, YU Tao, et al. Mixed game reinforcement learning of supply-demand interaction in power system dis-patch on electricity market. *Control Theory & Applications*, 2020, 37(4): 907 – 917.
(包涛, 李昊飞, 余涛, 等. 考虑市场因素的电力系统供需互动混合博弈强化学习算法. 控制理论与应用, 2020, 37(4): 907 – 917.)

[6] WANG Long, HUANG Feng. An interdisciplinary survey of multi-agent games, learning, and control. *Acta Automatica Sinica*, 2023, 49(3): 580 – 613, DOI: 10.16383/j.aas.c220680.
(王龙, 黄锋. 多智能体博弈、学习与控制. 自动化学报, 2023, 49(3): 580 – 613, DOI: 10.16383/j.aas.c220680.)

[7] LITTMAN M L. Markov games as a framework for multi-agent reinforcement learning. *Machine Learning Proceedings*. New Jersey: Elsevier, 1994: 157 – 163.

[8] SHAPLEY L S. Stochastic games. *Proceedings of the National Academy of Sciences*, 1953, 39(10): 1095 – 1100.

[9] LITTMAN M L, SZEPESVÁRI C. A generalized reinforcement–learning model: Convergence and applications. *Proceedings of the 13th International Conference on International Conference on Machine Learning*. San Francisco, CA, United States: Morgan Kaufmann Publishers Inc., 1996, 96: 310 – 318.

[10] WANG D, ZHAO M, HA M, et al. Stability and admissibility analysis for zero-sum games under general value iteration formulation. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, DOI: 10.1109/TNNLS.2022.3152268.

[11] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016, 529(7587): 484 – 489.

[12] MORAVČÍK M, SCHMID M, BURCH N, et al. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 2017, 356(6337): 508 – 513.

[13] BROWN N, SANDHOLM T. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 2018, 359(6374): 418 – 424.

[14] YANG Q, ZHANG J, SHI G, et al. Maneuver decision of UAV in short-range air combat based on deep reinforcement learning. *IEEE Access*, 2019, 8: 363 – 378.

[15] YOSHIDA S, ISHIHARA M, MIYAZAKI T, et al. Application of Monte-Carlo tree search in a fighting game AI. *The Global Conference on Consumer Electronics*. Kyoto, Japan: IEEE, 2016: 1 – 2.

[16] ZINKEVICH M, JOHANSON M, BOWLING M, et al. Regret minimization in games with incomplete information. *Proceedings of the 20th International Conference on Neural Information Processing Systems*. Red Hook, NY, United States: Curran Associates Inc., 2007: 1729 – 1736.

[17] ZHU Y, ZHAO D. Online minimax Q network learning for two-player zero-sum Markov games. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, 33(3): 1228 – 1241.

[18] YAZIDI A, SILVESTRE D, OOMMEN B J. Solving two-person zero-sum stochastic games with incomplete information using learning automata with artificial barriers. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, 34(2): 650 – 661.

[19] HARSANYI J C. Games with incomplete information played by "Bayesian" players, I–III Part I. The basic model. *Management Science*, 1967, 14(3): 159 – 182.

[20] FUDENBERG D, TIROLE J. *Game Theory*. Cambridge, MA, USA: MIT Press, 1991: 29.

[21] NASH JR J F. Equilibrium points in *n*-person games. *Proceedings of the National Academy of Sciences*, 1950, 36(1): 48 – 49.

[22] CHUA K, CALANDRA R, MCALLISTER R, et al. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Proceedings of the 31th International Conference on Neural Information Processing Systems*. Red Hook, NY, United States: Curran Associates Inc., 2018: 4759 – 4770.

[23] OSBORNE M J, RUBINSTEIN A. *A Course in Game Theory*. Cambridge, MA, USA: MIT Press, 1994: 13 – 14.

作者简介:

王成意 博士研究生, 目前研究方向为基于强化学习的非完全信息博弈, E-mail: wangchengyi@mail.ustc.edu.cn;

朱 进 副教授, 目前研究方向为随机博弈的理论及应用, E-mail: jinzhu@ustc.edu.cn;

赵云波 教授, 目前研究方向为人机混合智能和网络化智能控制, E-mail: ybzhao@ustc.edu.cn.