计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2014.20130823
51(10):2160-2170,2014

不完美信息扩展式博弈中在线虚拟遗憾最小化

胡裕靖[1] 高 阳[1] 安 波[2]
[1](软件新技术国家重点实验室(南京大学) 南京 210023)
[2](中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)
(huyujing.yujing.hu@gmail.com)

Online Counterfactual Regret Minimization in Repeated Imperfect Information Extensive Games

Hu Yujing[1], Gao Yang[1], and An Bo[2]
[1](*State Key Laboratory for Novel Software Technology* (*Nanjing University*), *Nanjing* 210023)
[2](*Key Laboratory of Intelligent Information Processing*, *Institute of Computing Technology*, *Chinese Academy of Science*, *Beijing* 100190)

Abstract In this paper, we consider the problem of exploiting suboptimal opponents in imperfect information extensive games. Most previous works use opponent modeling and find a best response to exploit the opponent. However, a potential drawback of such approach is that the best response may not be a real one, since the modeled strategy actually may not be the same as what the opponent plays. We try to solve this problem from the perspective of online regret minimization, which avoids opponent modeling. We make extensions to a state-of-the-art equilibrium-computing algorithm called counterfactual regret minimization (CFR). The core problem is how to compute the counterfactual values in online scenarios. We propose to learn approximations of these values from the results produced by the game and introduce two different estimators: static estimator which learns the values directly from the results' distribution, and dynamic estimator which assigns larger weight to new sampled results than older ones for better adapting to dynamic opponents. Two algorithms for online regret minimization are proposed based on the two estimators. We also give the conditions under which the values estimated by our estimators are equal to the true values, showing the relationship between CFR and our algorithms. Experimental results in one-card poker show that our algorithms not only perform the best when exploiting some weak opponents, but also outperform some state-of-the-art algorithms by achieving the highest win rate in matches with a few hands.

Key words extensive games; imperfect information; regret minimization; counterfactual regret minimization; static estimator; dynamic estimator

摘 要 研究在不完美信息扩展式博弈中对次优对手弱点的利用.针对该领域中一种常用方法——对手建模方法——的不足,提出了从遗憾最小化的角度来利用次优对手弱点的思想,并基于一种离线的均衡计算方法——虚拟遗憾最小化方法——将其扩展到在线博弈的场景中,实现对次优对手弱点的利用.提出了从博弈结果中估计各个信息集的虚拟价值的方法,给出2种估计手段:静态估计法和动态估计法.

收稿日期:2013-06-18;**修回日期**:2013-07-18
基金项目:国家自然科学基金项目(61035003,61175042,61321491,61202212);江苏省自然科学基金重点项目(BK2011005);江苏省普通高校研究生科研创新计划基金项目(CXLX13_049)
通信作者:高 阳(gaoy@nju.edu.cn)

静态估计法直接从博弈结果的分布中进行估计,并对每个结果给以相等的估计权重;而动态估计法则对新产生的博弈结果给以较高的估计权重,以便快速地适应对手的策略变化.基于2种估计方法,提出在线博弈中虚拟遗憾最小化的算法,并在基于单牌扑克的实验中,与4种在线学习算法(DBBR,MCCFR-os,Q-learning,Sarsa)进行了对比.实验结果显示所提出的算法不仅对较弱对手的利用效果最好,还能在与4种对比算法的比赛中取得最高的胜率.

关键词 扩展式博弈;不完美信息;遗憾最小化;虚拟遗憾最小化;静态估计法;动态估计法

中图法分类号 TP18

扩展式博弈(extensive games)是一种有效的顺序决策问题模型[1].许多问题都可以采用扩展式博弈来进行建模,比如国际象棋、围棋、扑克.近年来,许多研究者关注于扩展式博弈中纳什均衡(Nash equilibrium)的计算问题,并提出一些具有代表性的方法,包括状态约简和线性规划方法[2]、大间距技术(excessive gap)[3]、虚拟遗憾最小化(counterfactual regret minimization, CFR)[4]及其相关改进方法[5-7].

然而,当对手不使用纳什均衡策略进行博弈时(这样的对手称为次优对手),纳什均衡策略不能保证是对该对手的最佳反应(best response)策略.在这种情形下,利用(exploit)次优对手的弱点往往能够获得比使用纳什均衡策略更高的回报.在研究如何利用次优对手的问题中,绝大部分工作[8-9]采用了对手建模(opponent modeling)的方法,即首先对对手所使用的策略进行建模,然后对该策略模型计算出最佳反应.但是,由于对手建模所得出的策略模型不等同于对手所使用的真实策略,因此,针对策略模型的最佳反应与真正的最佳反应之间存在偏差.其次,对手建模的方法也无法对对手策略的动态变化作出快速反应.

为了避免对手建模,本文提出从遗憾最小化(regret minimization)的角度来利用次优对手的弱点.其原因是遗憾最小化关注于局中人(player)没有做什么、应该做什么,而非对手做了什么.基于一种离线的均衡计算方法——虚拟遗憾最小化CFR[4],本文提出了在线博弈中的虚拟遗憾最小化框架.在该框架中,核心问题是如何在线估计每个信息集(information set)的虚拟价值(counterfactual value, CF value).在CFR中,虚拟价值是基于对手的策略而定义的.由于在线博弈中对手的策略是未知的,本文提出从博弈的结果样本中来学习虚拟价值的近似值,并给出2种估计方法:静态估计法(static estimator)、动态估计法(dynamic estimator).静态估计法假设对手使用静态策略进行博弈,直接从博弈结果样本的分布中学习虚拟价值的近似值.相比之下,动态估计法则会对新产生的博弈结果赋予更高的权重,使虚拟价值能够体现对手策略的动态变化.基于这2种估计方法,本文分别提出了基于静态估计的在线虚拟遗憾最小化算法(static estimator-based OCFR, seOCFR)和基于动态估计的在线虚拟遗憾最小化算法(dynamic estimator-based OCFR, deOCFR).同时,也从理论上分析使用静态估计法和动态估计法所估计的虚拟价值与其真实值之间的关系.在基于单牌扑克(one-card poker)的实验中,与4种在线学习算法DBBR[9], MCCFR-os[7], Q-learning[10], Sarsa[10]进行了对比.实验结果显示本文所提出的算法不但能够更好地利用较弱的对手,还在与4种对比算法一对一的博弈中获得最高的胜率.

1 背景知识

1.1 扩展式博弈

与一般式博弈(normal-form game)不同,局中人(player)在扩展式博弈中必须按照一定的次序进行决策.根据信息的可见程度,扩展式博弈分为完美信息博弈和不完美信息博弈2种.本文仅对不完美信息情况下的扩展式博弈进行讨论,其定义如下:

定义1. 不完美信息扩展式博弈是一个六元组$\langle N,H,P,f_c,\{\mathfrak{T}_i\}_{i=1,2,\cdots,|N|},\{u_i\}_{i=1,2,\cdots,|N|}\rangle$,其中$N$是有限局中人(player)集合;$H$是有限动作序列(history)集合;空序列$\varnothing$是$H$中的元素;$H$中任意元素的每一个前缀也是$H$中的元素;终止序列(terminal history)是H中不属于其他任何序列前缀的序列,令$Z\subseteq H$为终止序列的集合;对于非终止序列$h\in H$,$A(h)=\{a|(h,a)\in H\}$表示在h上可以执行的动作集合;P是局中人指派函数;P从集合$N\cup\{c\}$为每一个非终止序列指派一个采取行动的局中人,其中c代表随机事件(chance);$P(h)$即表示在序列h之后将要执行动作的局中人;f_c是随

机事件概率函数;对于 $P(h)=c$ 的节点,$f_c(\cdot|h)$给定每一个 $A(h)$中的动作 a 的发生的概率,其值可用 $f_c(a|h)$来表示;对于局中人 $i\in N$,$\mathfrak{T}_i=\{I_i|I_i\subset H, P(I_i)=i\}$表示其信息分割(information partition);信息分割的元素称为信息集(information set),每个信息集是 H 的子集,代表若干无法明确区分的动作序列;一般地,对于任意 $h\in I_i$,可分别用 $A(I_i)$和 $P(I_i)$来代替 $A(h)$和 $P(h)$;对于每个局中人 $i\in N$,$u_i:Z\to\mathbb{R}$ 是其效用函数,规定其在每一个终止序列上所获得的效用值.

以上为不完美信息扩展式博弈的形式化定义.以单牌扑克为例,对扩展式博弈的定义作进一步解释说明.单牌扑克规则如下:

1) 2 个局中人,编号为 $P1$ 和 $P2$;

2) 一副牌包含 4 种花色的牌各 13 张,每种花色的点数从 2 到 A(2 最小,A 最大);

3) 开局时,2 个局中人首先分别投下 \$1 盲注;

4) 然后,每个局中人会随机发到 1 张扑克牌;

5) $P1$ 先采取动作:bet \$0(不下注,也叫过牌)或者 bet \$1(下注);

(1) 若 $P1$ 选择 bet \$0,$P2$ 可以选择 bet \$0 或 bet \$1;

①若 $P2$ 选择 bet \$0,点数高的局中人获得奖池中的 \$2;

②若 $P2$ 选择 bet \$1,$P1$ 可以 bet \$1 或 fold(弃牌);

Ⅰ. 若 $P1$ 选择 fold,则 $P2$ 获得奖池中的 \$3;

Ⅱ. 若 $P1$ 选择 bet \$1,点数高的局中人获得奖池中的 \$4;

(2)若 $P1$ 选择 bet \$1,$P2$ 可以 bet \$1 或 fold;

①若 $P2$ 选择 fold,则 $P1$ 获得奖池中的 \$3;

②若 $P2$ 选择 bet \$1,点数高的人获得奖池中的 \$4.

其博弈树可以如图 1 所示.在图 1 中,菱形代表随机事件节点(即发牌过程);方形表示 $P1$ 的决策节点;圆形表示 $P2$ 的决策节点;三角形节点表示博弈的结束.对应于扩展式博弈的形式化表示,可知单牌扑克中,局中人集合 $N=\{1,2\}$.现假设随机事件节点发给 $P1$ 的牌为♠K,给 $P2$ 发的牌为♥A.在这之后,若 $P1$ 首先选择动作 bet \$0,随后 $P2$ 也选择了 bet \$0,相应的历史动作序列可以表示为

h:〈c,♠K〉→〈c,♥A〉→〈1,\$0〉→〈2,\$0〉,

其中的每一项由相应的局中人和其选择的动作构成.然而,对于 $P1$ 来讲,它并不知道随机事件节点 c 给局中人 2 发的牌是♥A;同时,$P2$ 也不知道 $P1$ 得到了♠K.实际上,$P1$ 所观察到的是包含若干可能历史动作序列的信息集:

I_1:〈c,♠K〉→〈c,?〉→〈1,\$0〉→〈2,\$0〉,

$P2$ 观察到的信息集则为

I_2:〈c,?〉→〈c,♥A〉→〈1,\$0〉→〈2,\$0〉.

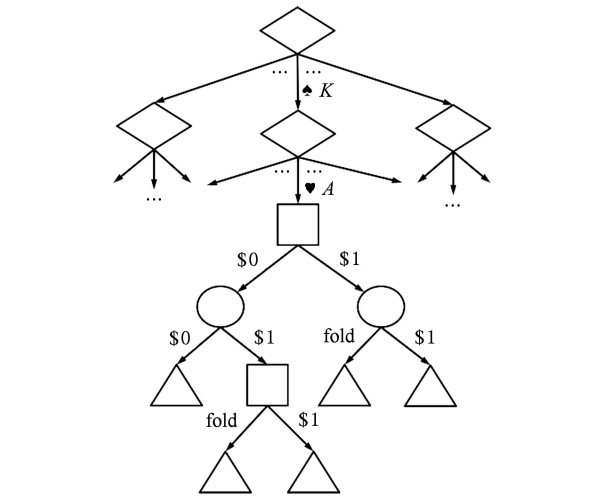


Fig. 1 The game tree of one-card poker.

图 1 单牌扑克的博弈树

1.2 策略的表示

对于局中人 $i\in N$,其策略可以表示为 σ_i.对于每一个信息集 $I_i\in\mathfrak{T}_i$,$\sigma_i(I_i):A(I_i)\to[0,1]$是在动作集 $A(I_i)$概率分布函数.局中人 i 的策略空间用 Σ_i 表示.一个策略组(strategy profile)包含所有局中人策略,用 $\sigma=(\sigma_1,\sigma_2,\cdots,\sigma_{|N|})$表示.一般地,对于局中人 i,我们用 σ_{-i}表示 σ 中除了 σ_i 之外的策略.

给定策略组 σ,动作序列 h 发生的概率表示为 $\pi^\sigma(h)$.显然,$\pi^\sigma(h)$可以分解为每一个局中人对其贡献的乘积:$\pi^\sigma(h)=\prod_{i\in N\cup\{c\}}\pi_i^\sigma(h)$.在这里,$\pi_i^\sigma(h)$即表示局中人 i 使用策略 σ_i 对 h 发生概率的贡献;同时,也可以定义除局中人 i 以外的其他局中人对 h 发生概率的贡献:$\pi_{-i}^\sigma(h)=\prod_{j\in N\cup\{c\}\setminus\{i\}}\pi_j^\sigma(h)$.

对于 2 个不同的动作序列 h 和 h',定义 $\pi^\sigma(h,h')$为在策略组 σ 下从 h 到 h'的转移概率.当 h 为 h'的前缀时,$\pi^\sigma(h,h')=\pi^\sigma(h')/\pi^\sigma(h)$;否则,该值为 0.类似地,也可以定义 $\pi_i^\sigma(h,h')$和 $\pi_{-i}^\sigma(h,h')$,这里不再赘述.局中人在给定策略组 σ 下所能得到的期望效用值可以如下计算:$u_i(\sigma)=\sum_{h\in Z}u_i(h)\pi^\sigma(h)$.

1.3 纳什均衡与遗憾最小化

定义 2. 最佳反应(best response):

局中人 i 对于所有其他局中人的策略组 σ_{-i}的

最佳反应策略 σ_i^* 满足:

$$u_i(\sigma_i^*,\sigma_{-i})\geqslant\max_{\sigma_i'\in\Sigma_i}u_i(\sigma_i',\sigma_{-i}). \quad (1)$$

在策略组 σ 中,若每一个局中人的策略都是其他所有局中人的策略组的最佳反应,那么 σ 就是一个纳什均衡(Nash equilibrium)策略.

定义 3. 纳什均衡.

策略组 $\sigma=(\sigma_1^*,\sigma_2^*,\cdots,\sigma_{|N|}^*)$ 是纳什均衡策略当且仅当对每一个局中人 $i\in N$,有:

$$u_i(\sigma)\geqslant\max_{\sigma_i'\in\Sigma_i}u_i(\sigma_1^*,\sigma_2^*,\cdots,\sigma_i',\cdots,\sigma_{|N|}^*). \quad (2)$$

在扩展式博弈中,通常较难计算精确的纳什均衡策略,而计算一个近似的均衡策略则比较容易.

定义 4. ε-纳什均衡

对于给定正实数 ε,策略组 σ 是 ε-纳什均衡当且仅当对于每一个局中人 $i\in N$,有:

$$u_i(\sigma)+\varepsilon\geqslant\max_{\sigma_i'\in\Sigma_i}u_i(\sigma_i',\sigma_{-i}). \quad (3)$$

假设扩展式博弈能重复地进行,令第 t 次博弈时的策略组为 σ^t. 若博弈已经进行 M 次,则这 M 次博弈中对于局中人 $i\in N$ 的平均遗憾值(average overall regret)定义为

$$R_i^M=\frac{1}{M}\max_{\sigma_i^*\in\Sigma_i}\sum_{t=1}^{M}(u_i(\sigma_i^*,\sigma_{-i}^t)-u_i(\sigma^t)). \quad (4)$$

关于纳什均衡与遗憾值之间的关系,有如下定理[11]:

定理 1. 在 2 人零和博弈中,在时刻 t,若 2 个局中人的平均遗憾值都小于 ε,那么 2 个局中人所使用的策略组的平均值 $\bar{\sigma}^t$ 是 2ε-纳什均衡策略.

1.4 CFR

CFR[4]是一种计算 2 人零和扩展式博弈中近似纳什均衡的有效方法,是一种反复迭代的过程,其理论基础是对各个信息集的虚拟遗憾值(counterfactual regret, CF regret)进行最小化,能够使得平均遗憾值最小化. 在每一次迭代中,所有信息集都会被访问到. 在每一个被访问的信息集节点上,CFR 计算出该信息集的虚拟价值和虚拟遗憾值(CF regret).

定义 5. 对于局中人 i 和信息集 $I\in\mathfrak{I}_i$,I 关于策略组 σ^t 的虚拟价值计算如下:

$$v_i(\sigma^t,I)=\sum_{z\in Z_I}u_i(z)\pi_{-i}^{\sigma^t}(z[I])\pi^{\sigma^t}(z[I],z), \quad (5)$$

其中,Z_I 是所有经过 I 的终止动作序列集合,$z[I]$ 代表终止序列 z 在 I 中的前缀.

定义 6. 对于局中人 i 和信息集 $I\in\mathfrak{I}_i$,每一个 I 上的动作 $a\in A(I)$ 在时刻 t 的虚拟遗憾值定义为

$$r_i^t(I,a)=v_i(\sigma_{I\to a}^t,I)-v_i(\sigma^t,I), \quad (6)$$

其中,$\sigma_{I\to a}^t$ 是一个与 σ^t 几乎相等的策略,但在信息集 I 上,前者总是选择执行动作 a. 其含义即是在信息集 I 上执行动作 a,相对于按照策略 σ^t 来采取动作的效用值之差. 若该值为正,则说明在信息集 I 上应该较多地执行动作 a,否则会产生遗憾.

CFR 每迭代一次之后,每个信息集上所得到的虚拟遗憾值都会被用来计算下一次迭代时的博弈策略. 新的策略通过遗憾值匹配(regret matching)来得到. 在这样反复迭代和策略更新的过程中,局中人的平均遗憾值被最小化,从而使得其平均策略向纳什均衡收敛.

除此之外还存在一些 CFR 的变体,例如:蒙特卡洛 CFR(MCCFR)[5-7]. 文献[7]将采样(sampling)机制引入到 CFR 的框架中,使得每一次迭代无需遍历整棵博弈树. 与 CFR 相比,尽管 MCCFR 需要更多的迭代次数,但其每一次迭代的时间却要少很多,同时也加速了策略的收敛;文献[6]在文献[7]的基础之上,对随机事件采样(chance sampling)方法进行了细化,提出了对手公共随机事件采样(opponent-public chance sampling)、自我公共随机事件采样(self-public chance sampling)、公共随机事件采样(public chance sampling)3 种方法,提高了计算纳什均衡策略的效率;而文献[5]则对蒙特卡洛 CFR 方法进行了理论上的概括,并提出了一种带试探的 CFR 方法(CFR sampling with probing),用以减小采样所带来的偏差. 国内对于扩展式博弈的研究大多集中在实际应用方面[12-13],目前尚无理论性的研究工作.

目前,基于 CFR 的相关工作所关注的问题都是如何对纳什均衡策略进行计算,并没有考虑在特定的对手情形下如何最大化博弈收益. 本文相对于之前工作的创新之处在于:1)将 CFR 方法扩展到在线博弈场景中,对次优对手的弱点进行利用,最大化博弈收益;2)将 CFR 与学习方法相结合,提出利用学习的机制来对信息集的虚拟价值进行估计.

2 在线虚拟遗憾最小化

本节基于遗憾最小化的观点,对 CFR 方法进行扩展,提出一种在扩展式博弈中利用次优对手的方法. 这其中的关键问题是,由于对手的策略未知,式

(5)所定义的虚拟价值无法在在线博弈的场景中直接进行计算.然而,实现遗憾最小化未必需要计算准确的虚拟价值.文献[5]中给出的理论结果表明:任意一种对虚拟价值的估计值都能够以一定的概率最小化遗憾值.基于该事实,本节提出从博弈结果中学习虚拟价值的近似值的思想.其原因在于博弈的结果是博弈双方策略的体现.基于这种思想,本节给出2种估计方法:静态估计法、动态估计法.静态估计法假设对手使用静态策略进行博弈,直接从博弈结果样本的分布中学习虚拟价值的近似值.相比之下,动态估计法则对新产生的博弈结果赋予更高的权重,使虚拟价值更能够体现对手策略的动态变化.

首先给出在线虚拟遗憾最小化的框架如下(该框架从局中人 i 的视角来描述):

算法 1. 在线虚拟遗憾最小化(online counterfactual regret minimization)框架.

步骤 1. 初始化.对于所有信息集 $I_i\in\mathfrak{T}_i$,其虚拟价值 $v_i(I_i)=0$;对于 I_i 上所有可执行的动作 $a\in A(I_i)$,其虚拟遗憾值 $R_i(I_i,a)=0$,其执行概率 $\sigma_i(I_i,a)=\dfrac{1}{|A(I_i)|}$.

步骤 2. 重复进行博弈.

for 每一局博弈:

① 按照 σ_i 与对手进行博弈;

② 记录博弈结果 u_i,并将其传递给虚拟价值估计器;

③ 对于每一个在该次博弈中被访问的信息集 I_i,通过虚拟价值估计器得到其近似虚拟价值 $v_i(I_i)$;

④ 对于每一个被访问的信息集 I_i,更新其虚拟遗憾值 R_i 和策略 σ_i;

end for

在每一次博弈中,局中人 i 根据当前策略 σ_i 与对手进行博弈,并得到相应的博弈结果.该结果将传递给虚拟价值估计器,由估计器计算出每一个被访问的信息集的虚拟价值,再更新其相应的虚拟遗憾值和策略,供下一次博弈使用.该框架并未交代具体的虚拟价值的估计方法以及更新虚拟遗憾值和策略的方法,它们将在算法 2 和算法 3 中给出.

2.1 静态估计法

事实上,在 CFR 中只有终止信息集(terminal information set)的虚拟价值才必须使用式(5)进行计算.对于一个非终止信息集 I,其虚拟价值能够通过其所有后继信息集的虚拟价值计算得出:

$$v_i(\sigma,I)=\sum_{I'\in Succ(I)}v_i(\sigma,I')\pi_i^\sigma(I,I'),$$

在这里,$Succ(I)$ 代表信息集 I 的后继信息集的集合,$\pi_i^\sigma(I,I')$ 为 I 到 I' 的转移概率.若在 I 中并非由局中人 i 采取动作,则 $\pi_i^\sigma(I,I')=1$.

所以,仅需要估计出各个终止信息集的虚拟价值,其他信息集的值可以通过以上方法进行计算.根据式(5),对于任意一个终止信息集 $I_z\in\mathfrak{T}_i$,静态估计法需要估计的值为

$$v_i(\sigma,I_z)=\sum_{z\in I_z}\pi_{-i}^\sigma(z)\pi_i^\sigma(I_z,z)u_i(z)=\sum_{z\in I_z}\pi_{-i}^\sigma(z)u_i(z).\tag{7}$$

静态估计法可以描述如下:

令 $M(\cdot)$ 表示一个信息集或终止动作序列被访问的次数.对于一个终止信息集 $I_z\in\mathfrak{T}_i$,$M(I_z)$ 可以计算如下:

$$M(I_z)=\sum_{z\in I_z}M(z).\tag{8}$$

若博弈在策略组 σ 不变的情况下重复进行多次,由大数定律可知,每一个终止序列 z 在信息集 I_z 上发生的次数将由 $\pi^\sigma(z)$ 决定.因此,对于 2 个不同的终止序列 $z_1,z_2\in I_z$,有 $\dfrac{M(z_1)}{M(z_2)}=\dfrac{\pi^\sigma(z_1)}{\pi^\sigma(z_2)}(M(z_2)\neq 0)$.又因为 $\pi_i^\sigma(z_1)=\pi_i^\sigma(z_2)$,则可以得到 $\dfrac{M(z_1)}{M(z_2)}=\dfrac{\pi_{-i}^\sigma(z_1)}{\pi_{-i}^\sigma(z_2)}$.对于任意终止序列 $z\in I_z$,重写式(7):

$$M(I_z)=\sum_{z'\in I_z}M(z')=\sum_{z'\in I_z}\frac{M(z)\pi_{-i}^\sigma(z')}{\pi_{-i}^\sigma(z)}=\frac{M(z)\sum_{z'\in I_z}\pi_{-i}^\sigma(z')}{\pi_{-i}^\sigma(z)}.$$

考察博弈树的初始节点 I_r(即博弈开始时所代表信息集),从 I_r 到某个终止信息集 I_z 的转移概率为 $\pi^\sigma(I_z)=\pi_i^\sigma(I_z)\sum_{z\in I_z}\pi_{-i}^\sigma(z)$.因此,$M(I_z)$ 与 $M(I_r)$ 之间有如下关系:

$$M(I_z)=M(I_r)\pi_i^\sigma(I_z)\sum_{z\in I_z}\pi_{-i}^\sigma(z).\tag{9}$$

在对 I_z 的 $M(I_z)$ 次访问中,记 $z^j\in I_z$ 为第 j 次访问时所发生的终止动作序列.显然,$M(I_r)$ 的值即博弈所发生的次数.定义局中人 i 在 $M(I_r)$ 次博弈中在 I_z 上所获得的平均效用值为

$$\hat{v}_i(\sigma,I_z)=\frac{\sum_{j=1}^{M(I_z)}u_i(z^j)}{M(I_r)}.\tag{10}$$

命题 1. $\hat{v}_i(\sigma,I_z)=\pi_i^{\sigma}(I_z)v_i(\sigma,I_z)$.

证明.

$\hat{v}_i(\sigma,I_z)$的定义为 $\hat{v}_i(\sigma,I_z)=\dfrac{\sum_{j=1}^{M(I_z)}u_i(z^j)}{M(I_r)}$,对于任意终止动作序列 $z\in I_z$,$M(I_z)$表示该动作序列在信息集 I_z 中出现的次数,所以 $\hat{v}_i(\sigma,I_z)$也可以写为

$$\hat{v}_i(\sigma,I_z)=\frac{\sum_{z\in I_z}u_i(z)M(z)}{M(I_r)}.$$

同时在分子分母上乘以 $M(I_z)$,可得:

$$\hat{v}_i(\sigma,I_z)=\frac{\sum_{z\in I_z}u_i(z)M(z)}{M(I_z)}\times\frac{M(I_z)}{M(I_r)}=\frac{M(I_z)}{M(I_r)}\times\sum_{z\in I_z}\frac{u_i(z)M(z)}{M(I_z)}. \tag{11}$$

当对信息集 I_z 的访问次数足够多时,根据大数定律,可以得到信息集 I_z 的访问次数 $M(I_z)$与任意终止动作序列 $z\in I_z$ 的访问次数 $M(z)$之间的关系:

$$M(I_z)=\sum_{z'\in I_z}M(z')=\sum_{z'\in I_z}\frac{M(z)\pi_{-i}^{\sigma}(z')}{\pi_{-i}^{\sigma}(z)}=\frac{M(z)\sum_{z'\in I_z}\pi_{-i}^{\sigma}(z')}{\pi_{-i}^{\sigma}(z)}. \tag{12}$$

将式(12)带入式(11)中,得到:

$$\hat{v}_i(\sigma,I_z)=\frac{M(I_z)}{M(I_r)}\times\sum_{z\in I_z}\frac{u_i(z)\pi_{-i}^{\sigma}(z)}{\sum_{z'\in I_z}\pi_{-i}^{\sigma}(z')}. \tag{13}$$

由式(9),可以得到 $M(I_z)$与 $M(I_r)$之间的关系:

$$\frac{M(I_z)}{M(I_r)}=\pi_i^{\sigma}(I_z)\sum_{z\in I_z}\pi_{-i}^{\sigma}(z). \tag{14}$$

将式(14)带入式(13)中,可得:

$$\hat{v}_i(\sigma,I_z)=\pi_i^{\sigma}(I_z)\sum_{z\in I_z}\pi_{-i}^{\sigma}(z)\sum_{z\in I_z}\frac{u_i(z)\pi_{-i}^{\sigma}(z)}{\sum_{z'\in I_z}\pi_{-i}^{\sigma}(z')}. \tag{15}$$

式(15)中分子和分母均有 $\sum_{z\in I_z}\pi_{-i}^{\sigma}(z)$这一项,所以简化为

$$\hat{v}_i(\sigma,I_z)=\pi_i^{\sigma}(I_z)\sum_{z\in I_z}u_i(z)\pi_{-i}^{\sigma}(z).$$

由于 $v_i(\sigma,I_z)=\sum_{z\in I_z}\pi_{-i}^{\sigma}(z)u_i(z)$,所以最终可以得到:

$$\hat{v}_i(\sigma,I_z)=\pi_i^{\sigma}(I_z)v_i(\sigma,I_z).$$

证毕.

因此,$\hat{v}_i(\sigma,I_z)/\pi_i^{\sigma}(I_z)$即为对信息集 I_z 的真实虚拟价值的估计($\pi_i^{\sigma}(I_z)\neq 0$). 从中可以看出,该估计值并没有使用对手的策略来进行计算. 当终止信息集的虚拟价值被估计出后,各个非终止信息集的虚拟价值将通过其后继的信息集的虚拟价值进行计算.

现给出基于静态估计法的在线虚拟遗憾最小化算法如下(以局中人 i 的视角描述):

算法 2. seOCFR 算法.

输入:当前博弈进行的次数 M、表 S_i、用于存储每一个终止信息集上的累积结果、存储每个信息集虚拟价值的表 v_i、存储每个信息集虚拟遗憾值的表 R_i.

输出:v_i,R_i,σ_i.

步骤 1. 进行一次博弈:

for 博弈的每一步:

① 根据当前策略 σ_i 进行博弈;

② if 某终止信息集 I_z 被访问

得到博弈结果 u_i;

$S_i(I_z)=S_i(I_z)+u_i$;

令 π_i^{σ} 为 σ_i 对到达 I_z 的贡献;

$v_i(I_z)=\dfrac{S_i(I_z)}{\pi_i^{\sigma}M}$;

end if

end for

步骤 2. 令 I_i 为在 I_z 之前被访问的信息集;

while $I_i\neq\varnothing$

① $v_i(I_i)=\sum_{I_i'\in Succ(I_i)}v_i(I_i')\pi_i^{\sigma}(I_i,I_i')$;

② if $P(I_i)==i$

for $A(I_i)$中的每一个动作 a

$R_i(I_i,a)=R_i(I_i,a)+v_i(Ia)-v_i(I)$;

end for

$\sigma_i(I_i)=regretMatching(I_i)$;

end if

令 I_i 为在其之前被访问的信息集;

end while

该算法即对应于算法 1 步骤 2 中 for 循环的内容. 在该算法中,步骤 1 完成对博弈结果的记录和终止信息集虚拟价值的更新,步骤 2 完成对所有被访问的非终止信息集的虚拟价值的更新以及策略的更新. 策略的更新采用遗憾值匹配的方式. 对于任意一个信息集 $I_i\in\mathfrak{T}_i$ 以及 $A(I_i)$中任意一个动作 a,遗憾值匹配[14]的方式如下:

$$\sigma(I_i,a)=\begin{cases}\dfrac{1}{|A(I_i)|}, & \sum_{a'\in A(I_i)}R_i^{+}(I_i,a')\leqslant 0,\\ \dfrac{R_i^{+}(I_i,a)}{\sum_{a'\in A(I_i)}R_i^{+}(I_i,a')}, & \text{otherwise},\end{cases}$$

其中,$R_i^+(I_i,a)=\max(R_i(I_i,a),0)$.

2.2 动态估计法

静态估计法实际上是利用了这样一个事实:如果局中人的策略保持不变,则博弈结果的分布即由双方的策略决定.然而,如果对手在博弈过程中不断地改变策略,博弈结果的分布就会不停地改变.在这种情况下,静态估计法未必有效.再次考查式(10),可以看出静态估计法实际上是为每个博弈结果分配了相同的大小为$\frac{1}{M(I_r)}$的权重.如果要更好地适应对手策略的动态变化,可以考虑为新产生的博弈结果分配较高的权重,而较早的博弈结果可以分配较低的权重.这就是动态估计法估计各个信息集虚拟价值的思想.

从方法上来讲,动态估计法借鉴了强化学习(reinforcement learning)的思想[10,15].强化学习的学习过程同CFR类似:首先估计状态或动作的效用值,再根据所估计的效用值更新策略.在动态估计法中,虚拟价值由强化学习所学得的动作效用值估计而来.其工作过程可以描述如下.

首先,根据强化学习的框架定义一些变量,如状态集合、值函数、奖赏函数等.对于局中人i,将其信息分割$\mathfrak{I}_i$作为强化学习中的状态集合,其所用策略仍表示为σ_i.对于任意信息集$I\in\mathfrak{I}_i$以及该信息集上的任意动作$a\in A(I)$,用$Q(I,a)$表示在I上采取动作a的效用值.仅考虑局中人i需要采取动作的信息集,因为只有在这些信息集上才能学习到Q值.从某个信息集转移到另一个信息集时,会产生相应的奖赏值(reward).假设局中人i在信息集I上执行了动作a,并转移到下一个需要采取动作的信息集I'上.若I'不是终止信息集,则从I到I'的奖赏值$r(I,a,I')=0$;否则,$r(I,a,I')=u_i(z)$,这里的z即是在I'上所观察到的终止动作序列.获取更多强化学习的信息[10,15].

对于信息集$I\in\mathfrak{I}_i$,定义其在策略组σ下的动态虚拟价值如下:

$$\tilde{v}_i(\sigma,I)=\frac{\sum_{a\in A(I)}Q(I,a)}{\pi_i^{\sigma}(I)}. \tag{16}$$

由于$Q(I,a)$是通过强化学习所学到的值,所以$\tilde{v}_i(\sigma,I)$的值不一定和真实的虚拟价值相等,但是二者却具有同样的含意:$Q(I,a)$代表在策略为σ的情况下,在信息集I上采取动作a的效用值;因此,$\sum_{a\in A(I)}Q(I,a)$即表示局中人i在I上使用策略σ_i的效用值;再由$\sum_{a\in A(I)}Q(I,a)$除以$\pi_i^{\sigma}(I)$得到$\tilde{v}_i(\sigma,I)$,表示去掉σ_i对于到达I的贡献.这与文献[4]中所解释的虚拟价值的含意是一致的.

基于以上过程,本文给出基于动态估计法的在线虚拟遗憾最小化算法如下:

算法 3. deOCFR算法.

输入:学习率$\alpha_1(0<\alpha_1<1)$和$\alpha_2(0<\alpha_2<1)$、存储每个信息集的动作效用值的表Q、存储每个信息集虚拟价值的表v_i、存储每个信息集虚拟遗憾值的表R_i.

输出:v_i,R_i,σ_i.

步骤 1. Q表各项初始化为0;

步骤 2. 令I_i为局中人i在博弈中遇到的第1个需要作决策的信息集;

步骤 3. 在I_i上根据策略σ_i选择一个动作a;

步骤 4.

for 博弈接下来的每一步

① 令I_i'为I_i之后需要作决策的信息集或终止信息集;

② 观察从I_i到I_i'的奖赏值r;

③ 在I_i'上根据策略σ_i选择一个动作a';

④ $Q(I_i,a)=\alpha_1Q(I_i,a)+\alpha_2(r+Q(I_i',a'))$;

⑤ $v_i(I_i)=\sum_{a\in A(I_i)}Q(I_i,a)$;

⑥ 令π_i^{σ}为σ_i对到达I_i的贡献;

⑦ for $A(I_i)$中的每一个动作a

$$R_i(I_i,a)=R_i(I_i,a)+\frac{Q(I_i,a)-v_i(I_i)}{\pi_i^{\sigma}}$$

end for

⑧ $\sigma_i(I_i)=regretMatching(I_i)$;

⑨ $I_i=I_i'$, $a=a'$;

end for

该算法的基本思想是利用强化学习的方法来学习各个信息集的虚拟价值.采用类似于文献[10]中的Sarsa算法的框架来学习各个信息集上的动作效用值,即下一个信息集I_i'上的动作效用值$Q_i(I_i',a')$被回退到信息集I_i,用于估计$Q(I_i,a)$的值.每一个博弈结果的在虚拟价值中的权重由学习率α_1和α_2决定.值得一提的是,在强化学习中通常规定$\alpha_1+\alpha_2=1$,在本算法中则没有此要求.

现给出动态估计法所估计出的值$\tilde{v}_i$与式(5)中所定义的真实值v_i之间的关系.考虑这样的信息集I_i:局中人在I_i采取动作后即到达一个终止信息集I_z.令t为目前访问I_z的次数,z^t表示当前在I_z中

所观察到的终止动作序列.由于在终止信息集上执行任何动作都没有意义,I_z 的 Q 值应为 0.对于 $Q(I_i,a)$,则有:

$$Q^t(I_i,a)=\alpha_1 Q^{t-1}(I_i,a)+\alpha_2 u_i(z^t),$$

其中,$Q^t(I_i,a)$代表 $Q(I_i,a)$在 I_i 第 t 次被访问后的值.令 $T_t(1\leqslant t\leqslant T_t)$表示在 I_z 第 t 次被访问时,博弈已经进行的次数.若将学习率 α_1 和 α_2 设置为可变的参数,令 $\alpha_1^t=\dfrac{T_{t-1}}{T_t}$,$\alpha_2^t=\dfrac{1}{T_t}$,则有:

$$Q^t(I_i,a)=\frac{T_{t-1}Q^{t-1}(I_i,a)+u_i(z^t)}{T_t}=$$

$$\frac{T_{t-1}\dfrac{T_{t-2}Q^{t-2}(I_i,a)+u_i(z^{t-1})}{T_{t-1}}+u_i(z^t)}{T_t}=$$

$$\frac{T_{t-2}Q^{t-2}(I_i,a)+u_i(z^{t-1})+u_i(z^t)}{T_t}=\cdots=$$

$$\frac{T_0Q^0(I_i,a)+\sum_{k=1}^{t}u_i(z^k)}{T_t}=\frac{\sum_{k=1}^{t}u_i(z^k)}{T_t}.$$

在以上推导中,最后一步是由于所有信息集的 Q 值均被初始化为 0.不难发现 T_t 即是在 2.1 节中所定义的 $M(I_r)$,所以有 $Q^t(I_i,a)=\hat{v}_i(\sigma,I_z)$.将 $Q^t(I_i,a)$带入式(11),可以得出:

$$\tilde{v}_i(\sigma,I_i)=\frac{\sum_{a\in A(I_i)}Q(I_i,a)}{\pi_i^\sigma(I_i)}=\frac{\sum_{I_z\in Succ(I_i)}\hat{v}_i(\sigma,I_z)}{\pi_i^\sigma(I_i)}=$$

$$\frac{\sum_{I_z\in Succ(I_i)}\pi_i^\sigma(I_z)v_i(\sigma,I_z)}{\pi_i^\sigma(I_i)}=$$

$$\frac{\sum_{I_z\in Succ(I_i)}\pi_i^\sigma(I_i)\pi_i^\sigma(I_i,I_z)v_i(\sigma,I_z)}{\pi_i^\sigma(I_i)}=v_i(\sigma,I_i).$$

由于以上推导中用到了静态估计法中的一些转换关系,所以这些等式只有在对手使用静态策略时成立.当对手使用动态策略时,更适宜将学习率 α_1 和 α_2 置为常量.如 2.2 节第一段所述,$\tilde{v}_i$ 实际上是关于博弈结果样本的加权和,而每一个博弈结果的权重则是若干学习率的乘积.当学习率为常量时,新产生的博弈结果的权重必定比较早的博弈结果的权重高,从而使得该算法能更好地适应对手策略的动态变化.

3 实验结果与分析

本文使用单牌扑克来评估所提出算法的性能,其规则已经在本文第 1 节给出.

首先测试了算法 seOCFR 和 deOCFR 利用较弱对手(weak opponent)的性能,并选择了 4 种算法来进行对比,包括权重转移的基于偏移的最佳反应算法(DBBR-WS)、基于结果采样的蒙特卡洛 CFR 算法(MCCFR-os)、Q-学习算法和 Sarsa 算法.DBBR-WS[9]是目前在线扩展式博弈中的最有效的算法之一,MCCFR-os 是文献[7]中对 MCCFR 算法的在线版本,而 Q-学习算法和 Sarsa 算法则是 2 种有效的强化学习算法.

3.1 实验 1

本文第 1 个实验将 seOCFR,deOCFR,DBBR-WS,MCCFR-os,Q-learning,Sarsa 算法作为 $P1$ 分别与 3 个不同类型的 $P2$ 进行博弈,包括随机对手(random opponent)、老练对手(sophisticated opponent)和动态对手(dynamic opponent).随机对手在每个信息集上根据固定的概率分布随机采取某个动作;老练对手在每个信息集上以 0.8 的采取随机动作,以 0.2 的概率按照纳什均衡采取动作;动态对手在 $\dfrac{1}{5}$的时间内采取随机动作,随后则会对 $P1$ 作出最佳反应.每一类 $P2$ 中包含 1 000 个不同的个体,每个个体在随机概率的选择上有不同.每一个作为 $P1$ 的算法与每一类 $P2$ 的所有个体分别进行比赛,每次比赛包含 20 000 局(hands).为减小实验误差,所有比赛的手牌都按照相同的规律发出(即前一次比赛的 20 000 局与后一次比赛的 20 000 局,每一局的牌都一样).图 2 中的 3 幅图显示了各个算法在与每一类对手进行博弈时,在比赛的每一局的胜率(平均胜率=当前赢得的总钱数÷当前完成的局数,且该胜率是对这一类对手的 1 000 个个体胜率的平均值).

从图 2 可以看出,与 CFR 相关的 3 种算法,即 seOCFR,deOCFR 和 MCCFR-os 相对于其他算法能够收敛到更高的胜率.由于随机对手和老练对手都使用静态策略,对 seOCFR 来说,其估计的虚拟价值与真实值几乎相等.所以 seOCFR 在对随机对手和老练对手的比赛中最终收敛的胜率比 deOCFR 和 MCCFR-os 稍高.另一方面,deOCFR 在对动态对手的比赛中获得了最高的胜率.其原因在于,deOCFR 对新产生的博弈结果赋予了更大的权重,使其能够较快地适应对手的策略变化.DBBR-WS 是一种对手建模方法,它首先对对手的策略进行建模,然后对对手采取最佳反应.然而,对对手策略所建立的模型不是对手真实使用的策略.当对手动态变换其策略

时,DBBR-WS 就无法快速适应. Q-学习算法和 Sarsa 算法在每个信息集上使用的是纯策略(即确定性的策略而非动作集上的概率分布),这在不完美信息扩展式博弈中未必奏效.

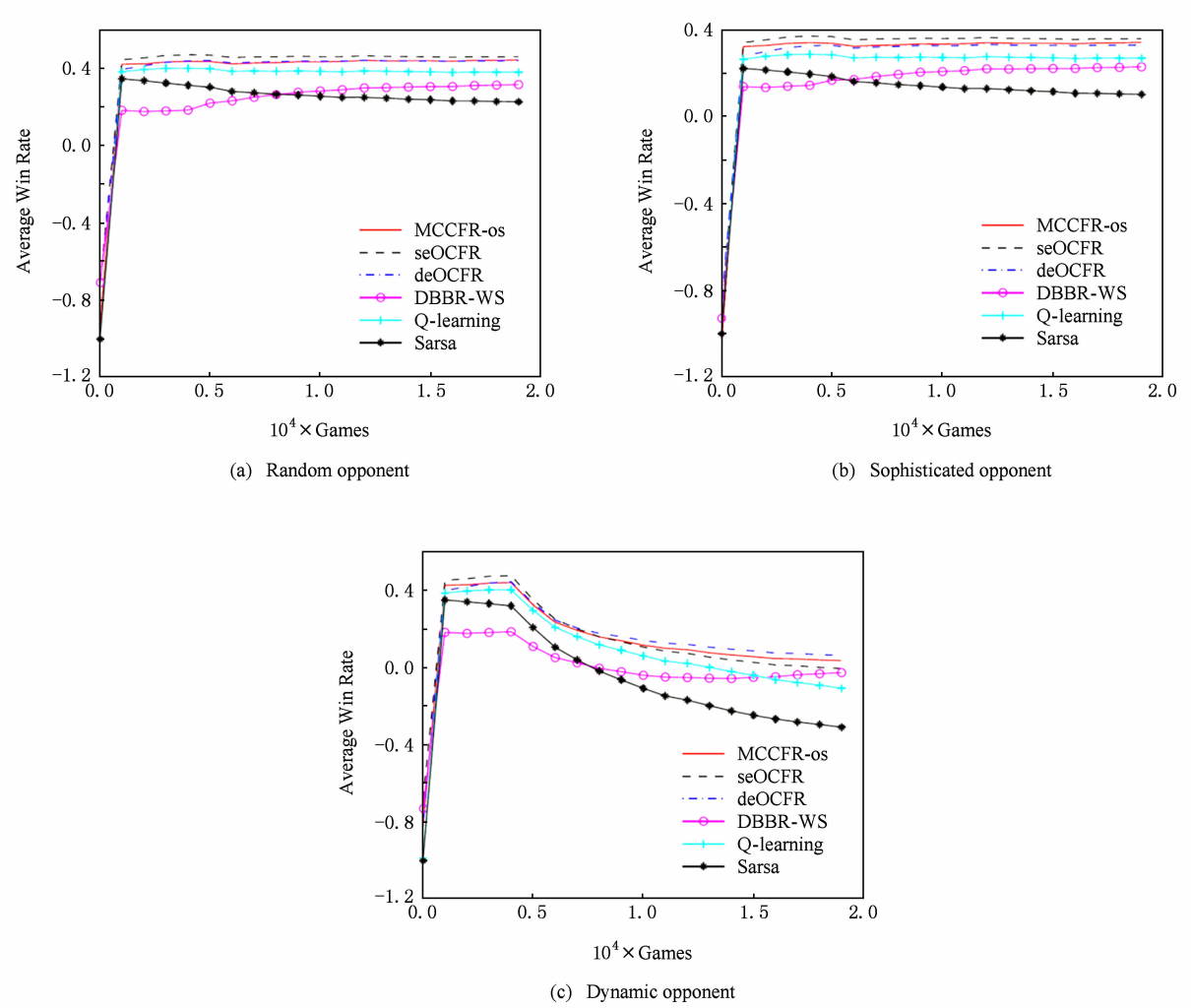


Fig. 2 Win rate curves of each tested algorithm in experiment 1.

图 2 实验 1 中各个算法的胜率曲线

3.2 实验 2

本文第 2 个实验测试 seOCFR 和 deOCFR 与其他学习算法博弈时的性能. 6 种算法相互之间进行比赛,每组比赛包含 1 500 局,每进行完一局之后 $P1$ 和 $P2$ 交换出牌位次. 记录比赛中每种算法的胜率,其结果如表 1 所示. 表 1 中,每一行代表某种算

Table 1 Win rates of each tested algorithm in one-on-one matches

表 1 各算法在一对一博弈中的胜率

Algorithms	DBBR-WS	MCCFR-os	Q-learning	Sarsa	seOCFR	deOCFR
DBBR-WS		−0.012±0.009	−0.005±0.0007	−0.004±0.0009	−0.007±0.0008	−0.038±0.0009
MCCFR-os	0.012±0.0009		0.053±0.0007	0.088±0.0008	−0.003±0.0007	−0.009±0.0007
Q-learning	0.005±0.0007	−0.053±0.0007		0.009±0.0007	−0.066±0.0007	−0.068±0.0006
Sarsa	0.004±0.0009	−0.088±0.0008	−0.009±0.0007		−0.094±0.0007	−0.098±0.0007
seOCFR	0.007±0.0008	0.003±0.0007	0.066±0.0007	0.094±0.0007		−0.005±0.0008
deOCFR	0.038±0.0009	0.009±0.0007	0.068±0.0006	0.098±0.0007	0.005±0.0008	

法与其他算法博弈时的胜率.可以看出,seOCFR 和 deOCFR 在与其他 4 种算法博弈时其胜率均为正.MCCFR-os 表现也相当出色,并且在与 DBBR-WS 的比赛中获得了比 seOCFR 高 0.005 的胜率.MCCFR-os 和 seOCFR 对虚拟价值进行估计时,都会利用每一次博弈的结果来对其重新进行计算,只是在计算方式上略有不同[6].其次,MCCFR-os 和 seOCFR 的虚拟遗憾值中均包含之前所有博弈结果的信息.所以,二者的博弈结果大致相当,仅有略微差别.对于 DBBR-WS,由于其对手都是学习算法,导致它无法建立一个准确的对手策略模型,使得它对其他所有算法的胜率均为负值.

4 结束语

本文关注于如何在不完美信息扩展式博弈中利用次优的对手来获得更高博弈收益的问题,并尝试从遗憾最小化的角度来解决.通过遗憾最小化,避免了对对手策略的建模.本文将一种均衡计算方法——虚拟遗憾最小化 CFR——扩展到在线博弈中,提出了从博弈的结果中估计信息集的虚拟价值的方法,并给出了 2 种估计手段:静态估计法和动态估计法.静态估计法假设对手使用静态策略,进而能直接从博弈结果的分布中估计虚拟价值;动态估计法则对新产生的博弈结果分配较高的权重,使其适应对手策略的动态变化.同时,也从理论上给出了 2 种估计方法所估计出的值与真实虚拟价值相等的条件.基于这 2 种估计方法,分别提出了 2 种在线虚拟遗憾最小化算法 seOCFR 和 deOCFR.在基于单牌扑克的实验中,本文所提出的 2 种算法与 4 种在线学习算法 DBBR,MCCFR-os,Q-learning,Sarsa 进行了对比.实验结果表明,seOCFR 和 deOCFR 不但能够较好地利用几类较弱的对手,也能在与 4 种对比算法一对一的比赛中获得最高的胜率.

参 考 文 献

[1] Osborne M, Rubinstein A. A Course in Game Theory [M]. Cambridge, MA: MIT Press, 1994: 200-201

[2] Billings D, Burch N, Davidson A, et al. Approximating game-theoretic optimal strategies for full-scale poker [C] // Proc of the 18th Int Joint Conf on Artificial Intelligence. Mahwah, NJ: Lawrence Erlbaum Associates, 2003: 661-668

[3] Hoda S, Gilpin A, Pena J, et al. Smoothing techniques for computing nash equilibria of sequential games [J]. Mathematics of Operations Research, 2010, 35(2): 494-512

[4] Zinkevich M, Johanson M, Bowling M, et al. Regret minimization in games with incomplete information [C] // Proc of the 21st Annual Conf on Neural Information Processing Systems. Vancouver, CA: Curran Associates Inc., 2007: 1729-1736

[5] Gibson R, Lanctot M, Burch N, et al. Generalized sampling and variance in counterfactual regret minimization [C] //Proc of the 26th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2012: 1355-1361

[6] Johanson M, Bard N, Lanctot M, et al. Efficient Nash equilibrium approximation through monte carlo counterfactual regret minimization [C] //Proc of the 11th Int Conf on Autonomous Agents and Multiagent Systems (AAMAS). Liverpool, UK: International Foundation of Autonomous Agents and Multi-Agent Systems (IFAAMAS) Press, 2012: 837-846

[7] Lanctot M, Waugh K, Zinkevich M, et al. Monte Carlo sampling for regret minimization in extensive games [C] // Proc of the 23rd Annual Conf on Neural Information Processing Systems. Vancouver, CA: Curran Associates Inc., 2009: 1078-1086

[8] Johanson M, and Bowling M. Data biased robust counter strategies [C] //Proc of the 12th Int Conf on Artificial Intelligence and Statistics (AISTATS). Brookline, MA: Microtome Publishing, 2009: 264-271

[9] Ganzfried S, and Sandholm T. Game theory-based opponent modeling in large imperfect-information games [C] //Proc of the 10th Int Conf on Autonomous Agents and Multi-Agent Systems (AAMAS). Liverpool, UK: International Foundation of Autonomous Agents and Multi-Agent Systems (IFAAMAS) Press, 2011: 533-540

[10] Sutton R, Barto A. Reinforcement Learning: An Introduction [M]. Cambridge, MA: MIT Press, 1998

[11] Hart S, Mas-Colell A. A simple adaptive procedure leading to correlated equilibrium [J]. Econometrica, 2003, 68(5): 1127-1150

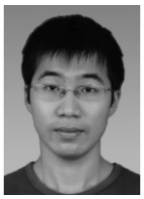
[12] Lin Dongmei. Extended rational secret sharing mechanism [J]. Computer Engineering, 2012, 38(16): 153-156 (in Chinese)
(林冬梅. 一种扩展的理性秘密分享机制[J]. 计算机工程, 2012, 38(16): 153-156)

[13] Zhong Liusheng, Cheng Lianglun. Unequal clustering energy-economical routing algorithm based on game-theory for WSN [J]. Application Research of Computers, 2009, 26(5): 1865-1867 (in Chinese)

(衷柳生, 程良伦. 基于博弈论的无线传感器网络非均匀分簇路由算法[J]. 计算机应用研究, 2009, 26(5): 1865-1867)

[14] Blackwell D. An analog of the minimax theorem for vector payoffs [J]. Pacific Journal of Mathematics, 1956, 6(1): 1-8

[15] Kaelbling L, Littman M, and Moore A. Reinforcement learning: A survey [J]. Journal of Artificial Intelligence Research, 1996, 4: 237-285



Hu Yujing, born in 1988. PhD candidate at the Department of Computer Science and Technology of Nanjing University. His main research interests include reinforcement learning, multi-agent learning, and game theory.

Gao Yang, born in 1972. PhD and professor at the Department of Computer Science and Technology of Nanjing University. Member of China Computer Federation. His main research interests include reinforcement learning, data mining, intelligent agent (gaoy@nju.edu.cn).

An Bo, born in 1980. PhD and associate professor at the Institute of Computing Technology of the Chinese Academy of Sciences. His research interests include artificial intelligence, multi-agent systems, game theory, automated negotiation, resource allocation, and optimization (boan@ict.ac.cn).