基于中介 Agent 的强化学习优化协商模型

张京敏 董红斌
(哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001)

摘 要 提出了一种基于强化学习的双边优化协商模型。引入了一个中介 Agent。在强化学习策略中使用不同的参数产生提议,进而选出最好的参数进行协商。为了进一步提高协商的性能,还提出了基于中介 Agent 自适应的学习能力。仿真实验结果证明了所提协商方法的有效性,且该方法提高了协商的性能。

关键词 多Agent,强化学习,自适应学习,中介 Agent
中图法分类号 TP18 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.01.010

Optimized Negotiation Model Based on Reinforcement Learning of Medium Agent

ZHANG Jing-min DONG Hong-bin
(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)

Abstract This paper proposed reinforcement learning bilateral optimized negotiation model based on reinforcement learning. The medium agent was introduced. It uses different parameters in the reinforcement learning strategy to produce proposals, and selects the best parameters to negotiate. The purpose is to further improve the performance of negotiation, and then the article presented the learning ability of adaptive based on medium agent. The simulation results show the effectiveness of the proposed method of negotiation and that it can improve the performance of negotiation.

Keywords Multi-agent system, Reinforcement learning, Adaptive learning, Medium agent

1 引言

协商是实现网上电子商务的重要手段,也是设计软件Agent的重要目标。如何提高 Agent 自主协商的能力一直是多 Agent 系统迫切需要解决的问题。

已经存在很多关于自主 Agent 协商的研究方法。比如:张化祥等人[4]把强化学习运用到谈判过程中,引入了价格信念、时间信念和时间贴现率。但是针对单属性问题,该算法存在让步快的缺点。孙天昊等人[5]采用的是贝叶斯分类的增强学习协商策略,该算法针对的是价格,而没有考虑到采用不同时间信念函数对协商的影响。张林兰等人[6]提出了基于自主 Agent 的并发多议题谈判框架,该框架采用的是一个封闭设计理念:基于 Agent 双方能够同时提交它们各自的提议给中介 Agent。该研究重要的贡献是协商议题不一定是对所有议题都是冲突的;在模型中考虑价格和数量的相关性,比较符合现实生活。孙天昊等人[7]采用的是对手分类的强化学习协商算法,该算法使 Agent 具有学习能力,针对对手的态度不断地进行改变,能够适应动态变化的环境,更好地达成协商,但是该方法是针对价格这个协商议题进行讨论的,模型相对简单。隋新等人[8]采用的是基于 Q-learing 的强化学习的 Agent 协商算法,仿真实验结果表明,该算法能够减少协商的时间,提高协商的效率,然而该模型中的多个议题是相互独立的。考虑到在现实生活中价格与数量有一定的必然联系,陈利红等人[9]提出了基于强化学习的双边多议题并行协商模型,该算法简单、稳定。但是文中的中介 Agent 起到协调者的作用,不仅仅是简单判断是不是存在交易机会、计算最终的协商协议,而且应该具备一定的学习能力,采用学习环境的知识,真正能够起到接近生活中的调解者的作用。现有的大量关于强化学习研究的协商模型中,大部分的学者研究的是协商模型,没有研究采用不同的时间信念函数对协商结果的影响;其次,不管是单议题,还是多议题相互独立的协商模型,在协商策略中使用强化学习都能够有效地提高协商的性能。与此同时,基于对手分类的强化学习算法充分利用了协商的历史信息,并根据对手的态度不断地调整协商的策略,能适应动态变化的环境,更能高效地达成谈判。

因此,本研究的目标是改进协商的方法,优化 Agent 的协商性能,提高中介 Agent 的自适应能力。主要贡献包含以下几部分:

1)考虑到强化学习在协商策略中存在一些很重要的参数,譬如时间信念知识、时间折扣率、协商轮次等,提出了基于强化学习双边优化协商模型,对不同的参数进行了对比,验证了时间信念为减函数、折扣率为 0.9 时,算法的性能更优。

2)考虑对手分类算法与 Q-learning 算法能够更好地适应动态变化的环境,本文提出了基于中介 Agent 自适应学习能力的对手分类算法,与强化学习双边优化协商算法比较,验证了该算法的可行性。

本文第 2 节基于强化学习双边优化协商模型,对模型的符号定义、效用评估、协商策略、协商算法进行了详细描述;第

到稿日期:2015-10-01 返修日期:2015-12-09 本文受国家自然科学基金项目(61472095,61272186),智能教育与信息工程黑龙江省高校重点实验室资助。

张京敏(1987-),女,硕士生,主要研究方向为机器学习、多 Agent 系统;董红斌(1963-),男,教授,博士生导师,主要研究方向为自然计算、机器学习、多 Agent 系统、数据挖掘,E-mail:donghongbin@hrbeu.edu.cn。

3 节介绍基于中介 Agent 自适应学习能力的对手分类协调协商，给出了中介 Agent 的自适应学习以及信念调整的过程；第 4 节给出实验结果，并对实验结果加以分析；最后对本文加以总结，并指出进一步的工作方向。

2 基于强化学习双边优化协商模型

这个模型协商关于议题价格和数量。引入了一个中介 Agent 作为协商机制，使用效用函数进行评估提议。

2.1 模型的符号定义

将协商模型定义为五元组：

$Ng=\langle G,A,D,U,T\rangle$

G：协商的主体，即协商参与者的 Agent 集合，定义 G 中包含 3 个子集 S，B 和 M 且 $S\cap B\cap M=\varnothing$，$S\cup B\cup M=G$，其中 S 表示卖方 Agent 集合，B 表示买方 Agent 集合，M 表示中介 Agent 集合。

A：协商议题集合，其中包含议题价格集合和议题数量。

D：协商议题的取值区间，$[RPs,IPs]$和$[IP_b,RP_b]$分别表示卖方 Agent 的价格区间和买方 Agent 的价格区间，在协商开始前这两个区间已知。

U：协商效用，进行评价协商结果，相应的效用函数将会在后面给出。其中 U 包含子集 U_b，Us，UR_b 和 URs，分别表示买方的效用值集合、卖方的效用值集合、买方的保留效用值集合以及卖方的保留效用值集合。保留效用值即是买卖双方能接受的最低效用值，也就是判断能否接受初始协议的标准。

T：协商截止时间，这里使用协商轮次表示协商时间。其中 T 包含子集 Ts 和 Tb 且 $T=\{0,1,2\cdots\}$，分别表示卖方的 Agent 的截止时间集合以及买方 Agent 的截止时间集合，截止时间即为双方 Agent 的最大协商轮次。

还有一些符号在给出相关的定义时再引出。

2.2 效用评估

(1)卖方 Agent

卖方 Agent 协商行为是基于它满意期望效用水平的。满意期望效用表示在某一轮次，卖方 Agent 能够承受的最小期望效用值。在谈判过程中，卖方 Agent 调整满意期望效用值采用的是强化学习算法。本节介绍卖方的期望效用评估。每个 Agent 都需要给予中介一些额外的费用 c。那么卖方的期望效用函数如下[6]：

$$U_s(p,q)=(p-RPs).q-c \tag{1}$$

可以看出卖方的期望效用函数是关于 p 和 q 的单调函数，而且都是增函数。其中 p 和 q 分别表示议题价格和数量。

(2)买方 Agent

买方 Agent 的购买数量 q 与市场的需求 x 有关。假定 x 分布在区间$[a,b]$，那么购买数量 q 也分布在区间$[a,b]$。p_0 是出售价格，如果 $q>x$，买方 Agent 将以低的价格 αp_0 卖出商品，α 是折扣率。一般情况下，αp_0 比购买价格 p 低，出售的价格 p_0 比购买价格 p 高，因此 αp_0 小于 IP_b，p_0 大于 RP_b，$p\in[IP_b,RP_b]$。给出一个协商议题$\langle p,q\rangle$，买方的利润函数(期望效用函数)f 如下所示：

$$f(p,q,x)=\begin{cases}(p_0-p)\cdot q, & \text{if } q\leqslant x\leqslant b\\ (p_0-p)\cdot x+(\alpha p_0-p)\cdot(q-x), & \text{if } a\leqslant x<q\end{cases}$$

可以看出利润函数是关于变量 x 的函数。所以，对于一个协商提议$\langle p,q\rangle$来说，买方的期望效用函数如下所示：

$$\begin{aligned}U_b(p,q)&=\int_a^b f(p,q).\frac{1}{b-a}\mathrm{d}x-c=\frac{1}{b-a}\Big(\int_a^q(p_0-p)\cdot x+(\alpha p_0-p)\cdot(q-x)\,\mathrm{d}x+\int_q^b(p_0-p)\cdot q\,\mathrm{d}x\Big)-c\\&=\frac{1}{b-a}\cdot\Big(-\frac{(1-\alpha)p_0}{2}\cdot q^2+(p_0(b-\alpha a)-p(b-a))\cdot q-\frac{(1-\alpha)p_0}{2}\cdot a^2\Big)-c\end{aligned} \tag{2}$$

(3)中介 Agent

在这个模型中，中介 Agent 的任务是收到双方的提议$\langle p,q\rangle$后，判断是不是存在交易机会，如果不存在交易的条件，那么通知双方进入下一轮，分别提交一个新的协商议题；如果存在交易的条件，那么中介 Agent 在下一轮提交一个可能数量范围内的所有提议，计算出同一数量 q 下价格差最大的两个值，那么最终的协商价格是两个价格的平均值，协商的数量是价格差最大时对应的数 q。

2.3 强化学习协商策略

中介 Agent 同时接收买卖双方 Agent 的协商提议，在协商未成功之前，买卖双方的回报值为 0。协商成功之后的回报值为 r，即前面定义的期望效用。就是说在最后达成协商之后，买卖双方 Agent 才能够收到相应回报 r。假设达成了协商，谈判议题的数量为 q^T，协商议题价格为 p^T，那么买卖双方 Agent 回报如下：

卖方：$r_s=U_s(p^T,q^T)=(p^T-RPs)\cdot q^T-c$

买方：$r_b=U_b(p^T,q^T)$

Agent 在当前状态下选择最优动作转化到下一状态的过程是 Agent 产生协商议题的过程。举例：当协商轮次是 t，Agent状态为 $s(t)$时，处理完协商议题$\langle p(t),q(t)\rangle$时，状态变化为 $s(t+1)$。

基于 Q-学习的 Q 函数定义为：

$$Q(s(t),p(t),q(t))=r(s(t),p(t),q(t))+\gamma\max_{p(t+1),q(t+1)}Q(\delta(s(t),p(t),q(t)),p(t+1),q(t+1)) \tag{3}$$

其中，时间贴现率为 γ，当前的状态为 $s(t)$下处理完协商议题$\langle p(t),q(t)\rangle$时产生的立即奖赏值为 $r(s(t),p(t),q(t))$，状态转移函数为 $\delta()$。当协商取得成功时，买卖双方 Agent 才能够获得相应的回报值。在协商的过程中回报值是 0，达成了协商时回报值为正，协商失败时回报值为负[10]。因此，买卖双方 Agent 都抱有最大的诚意达成协商，以获得较高的回报值。

若第 t 次提交协商议题获得成功，取得的回报值是 Qe。达成协商的轮次与每次提交协商议题轮次进行区分，第 t 次时提交协商议题的轮次称作第 t 阶段进行提议。因此，由 Q-learning 中 Q 函数定义：

若第一次提交协商议题$\langle p(1),q(1)\rangle$达成协商时，第一阶段 Q 函数值为 $Q(s(1),p(1),q(1))=Qe$；

第二次提交协商议题$\langle p(2),q(2)\rangle$达成协商时，第一阶段 Q 函数值为 $Q(s(2),p(2),q(2))=\gamma Qe$；

…

第 t 次提交议协商议题$\langle p(t),q(t)\rangle$达成协商时，第 t 阶段 Q 函数值为 $Q(s(t),p(t),q(t))=r^{t-1}Qe$。

所以第一阶段卖方平均回报值如下：

$$\overline{Qse}(s(1),p(1),q(1))=\frac{\sum_{i=1}^{Ts}sfb(i)\gamma^{i-1}Qse}{Ts} \tag{4}$$

推导可得，第 t 阶段卖房的平均回报值为：

$$\overline{Qse}(s(t),p(t),q(t))=\frac{\sum_{i=t}^{Ts} sfb(i)\gamma^{i-t}Qse}{Ts-t+1} \tag{5}$$

第 t 阶段买方的平均回报值为：

$$\overline{Qbe}(s(t),p(t),q(t))=\frac{\sum_{i=t}^{Tb} bfs(i)\gamma^{i-t}Qbe}{Tb-t+1} \tag{6}$$

由平均回报函数可知，它是关于时间信念的函数，在已有的文献中时间信念函数 $sfb(i)$ 采用固定的函数，本文提出买方和卖方对时间信念的增函数、减函数和常时间函数进行研究。观察采用不同的时间信念函数以及时间贴现率对协商过程的影响。

由前面介绍的期望效用函数，能够看出买卖双方的期望效用是关于价格 p 和数量 q 的二元函数。所以，达成协商时买卖双方的回报值是价格和数量在定义域内的二重积分。在第 t 协商轮次达成协商时，买卖双方 Agent 的回报值如下。

卖方 Agent 回报值：

$$\begin{aligned}Qse&=\iint_D r_s\cdot spb\cdot sqb\cdot \mathrm{d}p^{\mathrm{T}}\mathrm{d}q^{\mathrm{T}}\\&=\int_a^b\int_{RPs}^{IPs} r_s\cdot\frac{1}{IPs-RPs}\cdot\frac{1}{b-a}\cdot\mathrm{d}p^{\mathrm{T}}\mathrm{d}q^{\mathrm{T}}\end{aligned} \tag{7}$$

买方 Agent 回报值：

$$\begin{aligned}Qbe&=\iint_D r_b\cdot bps\cdot bqs\cdot \mathrm{d}p^{\mathrm{T}}\mathrm{d}q^{\mathrm{T}}\\&=\int_a^b\int_{IPb}^{p_0} r_b\cdot\frac{1}{p_0-IPb}\cdot\frac{1}{b-a}\cdot\mathrm{d}p^{\mathrm{T}}\mathrm{d}q^{\mathrm{T}}\end{aligned} \tag{8}$$

最后，卖方 Agent 的报价策略为：

$$p_s(t)=\frac{\lambda_s\overline{Qse}+c}{q_s(t)}+RPs \tag{9}$$

买方 Agent 的报价策略为：

$$p_b(t)=\frac{b-\alpha a}{b-a}\cdot p_0-\frac{(1-\alpha)p_0((q_b(t))^2+a^2)}{2(b-a)q_b(t)}-\frac{\lambda_b\overline{Qbe}+c}{q_b(t)} \tag{10}$$

强化学习协商过程中由于妥协过快的缺点，如式(9)和式(10)引入了期望还原率 λ，λ 是有限制的，不能够取负数，因为期望效用只能是正数，因此 λ 最小值是 0。求得 λ 的最大值，根据强化学习的协商策略，能够知道 Q 值就是期望效用，买卖双方 Agent 由期望效用公式产生提议 $\langle p(t),q(t)\rangle$，因此 $\lambda\overline{Qe}(t)$ 的最大值就是期望效用的最大取值 $U_{\max}$。根据下面定义 1 和定义 2 可分别求出卖方和买方的最大效用，所以 $\lambda_{\max}=U_{\max}/\overline{Qe}(t)$。根据式(5)和式(6)可知，$\overline{Qe}(t)$ 是一个随着 t 的增大而逐渐变小的值，事实上当 $t=T$ 时，λ 值最大，但是 $\lambda\overline{Qe}(1)=U_{\max}\overline{Qe}(1)/\overline{Qe}(T)>U_{\max}$，这是不正确的，所以 λ 的最大值 $\lambda_{\max}=U_{\max}/\overline{Qe}(1)$。综上，$\lambda$ 的范围是 $[0,\lambda_{\max}]$，其中，$\lambda_{\max}=U_{\max}/\overline{Qe}(1)$。

考虑到协商的过程中更加趋近于现实生活，随着时间逐渐增大的过程，双方的期望效用逐渐降低，让步的程度逐渐增大。本文采用双方的期望还原率 λ 均为 $\beta(t)\lambda_{\max}$，其中，$\beta(t)=1-t/T$，$\lambda_{\max}=U_{\max}/\overline{Qe}(1)$，$U_{\max}$ 为最大的期望效用[11]。

从买方和卖方 Agent 的报价公式可以看出价格是关于数量的一次函数，因此必须先求出数量值才能够得到价格的值，得到数量值后由参考文献[6]给出了价格与数量的定义。

定义 1　在协商初始阶段，卖方 Agent 的最佳协商议题 $\langle p_s(0),q_s(0)\rangle$ 符合下面的公式：

$p_s(0)=IPs, q_s(0)=b$

由于在开始阶段采用的策略使期望效用值最大，随着协商过程逐渐进行，期望效用不断减少。期望效用函数是关于价格和数量的单调增函数，因此数量取最大值，价格取最大值，即 $p_s(0)=IPs, q_s(0)=b$。

定义 2　在协商初始阶段，买方 Agent 的最佳的协商议题 $\langle p_b(0),q_b(0)\rangle$ 符合下面的公式：

$p_b(0)=IPb$

$q_b(0)=Ceiling(((b-\alpha a)p_0-(b-a)p_b(0)/((1-\alpha)p_0))$

在协商的开始阶段，为了使买方的期望效用最大化，由买方的期望效用函数就能够求出

$p_b(0)=IPb$

$q_b(0)=Ceiling(((b-\alpha a)p_0-(b-a)p_b(0)/((1-\alpha)p_0))$

定义 3　在协商第 $t(t\geqslant 1)$ 阶段时，卖方 Agent 的最佳协商议题 $\langle p_s(t),q_s(t)\rangle$ 满足下面的公式：

$$p_s(t)=\frac{\lambda_s\overline{Qse}+c}{q_s(t)}+RPs, q_s(t)=b$$

数量总是设定为最大值，采用市场上薄利多销的原则，价格是由 Q-learning 算法取得。

定义 4　在协商第 $t(t\geqslant 1)$ 阶段时，买者 Agent 的最佳的协商议题 $\langle p_b(t),q_b(t)\rangle$ 满足下面的公式：

$$p_b(t)=\frac{b-\alpha a}{b-a}\cdot p_0-\frac{(1-\alpha)p_0((q_b(t))^2+a^2)}{2(b-a)q_b(t)}-\frac{\lambda_b\overline{Qbe}+c}{q_b(t)}$$

$$q_b(t)=Ceiling(\sqrt{a^2+2(b-a)(\lambda_b\overline{Qbe}+c)/((1-\alpha)p_0)}).$$

数量通过价格关于数量的一阶导数进行计算求出。

2.4　协商算法

我们定义的算法是依据下面的假设情况：卖方 Agent 有一个效用期望函数满足式(1)，买方 Agent 有一个效用期望函数满足式(2)，每一个的报价过程采用的是强化学习中的 Q-learning算法。双方的 Agent 在一定的协商范围内进行讨价还价，根据下面的步骤：

1)卖方 Agent 和买方 Agent 很明确描述出协商价格以及数量取值范围，并把数量的取值范围发送给中介 Agent，中介 Agent 根据市场的分析和提交的数量范围，最终确定数量 q 的范围为 $[a,b]$，并发送给双方 Agent。

2)卖方 Agent 和买方 Agent 发送初始协商议题给中介 Agent，在协商的初始阶段，卖方 Agent 依据定义 1 产生初始谈判议题，同样买方 Agent 依据定义 2 产生初始谈判议题。

3)在之后每一轮 $t(t\geqslant 1)$，卖方 Agent 和买方 Agent 依据定义 3 和定义 4 产生他们各自的协商议题。

4)卖方 Agent 每次报价都会保证谈判的数量是最大的，总会有 $q_s>q_b$，所以，中介 Agent 接收来自双方 Agent 的谈判议题以后，如果存在 $p_b>p_s$，双方就存在交易机会，否则就不存在交易机会。

若在 T 轮，中介 Agent 确定有一个交易机会，它会通知各个 Agent 在下一轮即 $T+1$ 轮提交数量 $[q_s(T)q_b(T)]$ 在所有的谈判议题对。买方产生的议题对的效用值都是相等的，卖方产生的议题对的效用值也都是相等的。注意，$T+1$ 轮和 T 轮的效用是相等的。关于相同的数量 q 下所有的谈判议题对不存在 $p_b(T+1)>p_s(T+1)$，那么，中介 Agent 会通知各

个 Agent 进入下一轮的谈判。如果关于相同的数量 q 下所有的谈判议题对存在 $p_b(T+1)>p_s(T+1)$，中介 Agent 会选择同一个数量的前提下 $p_b(T+1)-p_s(T+1)$ 的最大值的谈判对作为最终达成谈判的议题。最终谈判议题的价格取 $(p_b(T+1)+p_s(T+1))/2$，而数量取 $p_b(T+1)-p_s(T+1)$ 获得最大值条件下数量。

5)当中介 Agent 把最终达成的协商议题发送给双方 Agent时，双方计算各自的效用值，若得到双方的效用值都大于各自的保留效用值时，那么就取得最终的协商；相反，双方 Agent 把拒绝该提议的消息发送给中介 Agent。还没达到截止时间时，中介 Agent 就会通知进行下一轮协商。

6)协商的过程一直进行，直到达成协商或者某一方到达截止时间退出协商。

算法的流程图如图 1 所示。

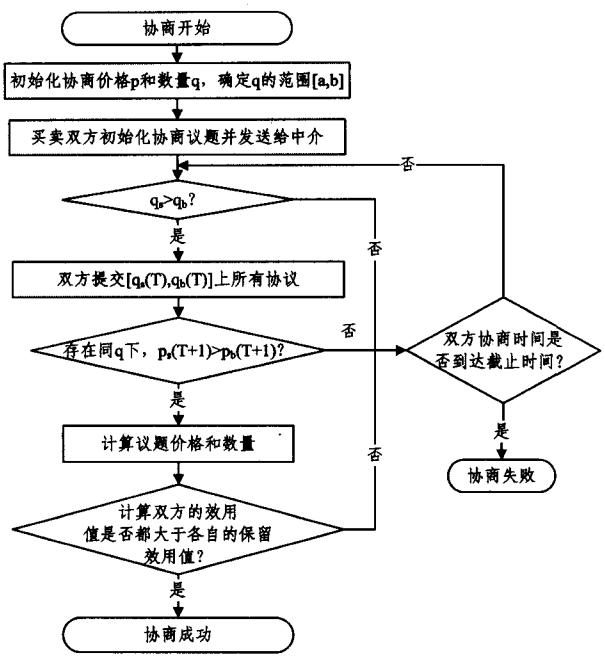

图 1 算法流程图

3 基于中介 Agent 自适应学习能力的对手分类协调协商

目的是进一步提高协商性能和优化中介 Agent 的调节能力。提出一个基于中介 Agent 自适应学习能力的对手分类协调协商算法。为了符合现实生活中人类的协商心理，在强化学习的时间信念函数为减函数、时间折扣率为 0.9 进行提议的基础上，介绍了一个对手分类算法进一步地优化协商的性能，使得 Agent 在动态变化的环境中能够根据对方的信念进行推理，更新自身的信念，自主地提高协商的能力，有效地与对手进行协商，使得利益最大化。

3.1 中介 Agent 的自适应学习

在 Agent 的自动协商过程中，在每一轮协商过程中双方都会有协商的历史信息，那么应该如何充分利用这些信息，并加快协商的过程，从而提高协商的解？本节就是介绍如何利用对手报价的历史信息。

所谓对手行为的谈判策略就是利用谈判的历史信息来评价对手的态度，通过对手的态度对对手进行分类，再根据某一种算法确定自己这方的协商议题值和妥协程度。因为采用对手的态度不断变化，自己采取相应的协商策略，所以是适应动态变化的环境，能够更好地达成协商。

本文介绍的是数量和价格相关的多议题协商模型，这里为了方便及易理解地介绍理论知识，先给出一个单议题的例子，其与本节介绍的多议题在理论上是没区别的。

议题为 x，协商的参与者为 a 和 b，在协商过程中 Agent a 的提议顺序就是协商的历史信息 H_a^t，该顺序如下：

$$x_{a\to b}^1, x_{a\to b}^2, x_{a\to b}^3, \cdots, x_{a\to b}^t, 1\leqslant t_1<t_2<\cdots<t<t_{\max}^a$$

其中，t 为当前时刻。

按照让步程度的不同，将其分为 4 类：绝对平均让步、绝对最小让步、绝对最大让步、相对平均让步。

(1)绝对平均让步

$$\Delta_{AACD}^a=(last(H_a^t)-first(H_a^t))/len(H_a^t) \tag{11}$$

谈判的历史信息中首个提议为 $first(H_a^t)$；

谈判的历史信息中最后提议为 $last(H_a^t)$；

谈判的历史信息的总长度为 $len(H_a^t)$。

(2)绝对最小让步

$$\Delta_{AMinCD}^a=\min(x_{a\to b}^t-x_{a\to b}^{t-2}), 2<t\leqslant t_{\max}^a \tag{12}$$

(3)绝对最小让步

$$\Delta_{AMaxCD}^a=man(x_{a\to b}^t-x_{a\to b}^{t-2}) \tag{13}$$

(4)相对平均让步

$$\Delta_{RACD}^a=man(x_{a\to b}^t-x_{a\to b}^{t-2})/k \tag{14}$$

在协商的历史信息 H_a^t 中，基于某一个时间点比如 t 点，那么 t 点之前的 k 个协商议题的平均让步程度是：

$$1\leqslant t-2k<t\leqslant t_{\max}^a, k=1,2,\cdots$$

采用这个决策时，Agent a 下一次也就是第 $t+1$ 次的协商议题值为：

$$x_{a\to b}^{t+1}=\min(\max(x_{a\to b}^{t-1}+f_j^{behavior}(H_b^t), low_j^a, high_j^a)) \tag{15}$$

让步函数为：

$$f_j^{behavior}(H_b^t)\in\{\Delta_{AACD}^a, \Delta_{AMinCD}^a, \Delta_{AMaxCD}^a, \Delta_{RACD}^a\} \tag{16}$$

H_b^t 为 Agent 在协商过程中的协商记录，如果要对对手进行分类，那么 $len(H_b^t)>3$。

在 Q-learning 算法与对手分类进行结合时，根据前面让步函数的详细介绍，对对手的协商历史信息进行学习，得出对手属于哪种类型，对对手的信念知识进行动态的改变，最后通过不同的让步策略与对手协商。

定义 5 在谈判过程中，卖方 Agent 的议题序列 $H_s(t)$（卖方 Agent 的谈判历史信息）：

$$H_s(t)=p_s(0)p_s(1)\cdots p_s(t)\ (t>3)$$

其中，$p_s(0)=p_s^{\max}$。买方 Agent 的议题序列定义类似，但其中 $p_b(0)=p_b^{\min}$。

在谈判(协商)历史信息中，谈判对手所表现出来的特征分为 3 种类型：让步型(C)、执着型(O)、均匀线型(L)。

本节使用比较常用的绝对平均让步函数，比如卖方 Agent，即 $\Delta=(p_s^t(t)-p_s^0)/t, t>0$。绝对平均让步函数与谈判初始值的比叫作让步比例，记作 $\alpha=\Delta/p_s^{\max}$。那么，下面就是对手类型：

$$C=\begin{cases}C_C, & \alpha>\theta_1\\ C_L, & \theta_2\leqslant\alpha\leqslant\theta_1\\ C_O, & \alpha<\theta_2\end{cases} \tag{17}$$

进行分类的上限和下限是 θ_1 和 θ_2，比如 $\theta_1=0.1, \theta_2=0.05$。

3.2 信念调整过程

对手的类型分类依据对手的让步程度，所以采用哪种协

商策略是通过在协商的过程中根据不同的对手类型进行判断。在强化学习算法中的时间信念函数也就是谈判的态度，t/T 表示的是执着类型的，$1-t/T$ 表示的让步型的。那么把强化学习算法与对手的分类进行组合，判断对手的类型，从而改变时间信念函数，提高协商的解的质量。原则为：

1)假设对手是执着类型，那么时间信念函数改变为减函数；

2)假设对手为让步型，那么时间信念函数改变为增函数(t/T)；

3)假设对手为均匀线型，那么时间信念函数改变为常函数。

以买家为例：

$$b^t(t)=\begin{cases} t/T_b, & C^s=C_C \\ 1-t/T_b, & C^s=C_O \\ 0.5, & C^s=C_L \end{cases} \tag{18}$$

上面介绍的双方协商是关于单议题的，那么把这对手分类的思想运用到基于强化学习双边优化协商模型中，使中介 Agent 具备自适应学习能力。

进一步优化的算法与原始模型的算法大致相同，仅第 4 步进行改变，变动的部分如下。

中介 Agent 存储协商过程的卖方的每次报价，根据协商历史信息算出卖方的让步比例 α；然后根据 α 确定本次协商议题时卖方属于哪种类型，中介 Agent 告知买方对手卖方的类型，在得到卖方类型后改变时间信念函数，Q-learning 算法将根据新的时间信念函数来算出下次协商议题(价格和数量)值。后面部分与原始算法第 4 步完全相同，这里不再介绍。

4　实验的设计与分析

目的是证明所提算法的有效性和高效性，对比实验在 VC++平台上进行，一个是对强化学习双边优化协商中的重要参数做的实验，另一个是对中介 Agent 自适应学习能力的对手分类算法进行的实验。

表 1　参数设置

Parameters	Buyer	Seller
RP	—	[83 75 72 60]
IP	[68 62 60 55]	[110 100 97 86]
UR	[605.5 655.5 665.5 712.5]	[270 240 230 200]
Td	12	10
α	0.5	—
p_0	100	—
A	100	
B	150	
Γ	0.9	

首先，初始化一些重要的参数，表 1 列出实验所需的参数以及赋值。考虑一个卖方 Agent S 和一个买方 Agent B 协商关于一个特定产品。两个协商者针对议题价格 p、数量 q 进行谈判。Agent S 和 Agent B 需要通过中介 Agent M 进行协商，所以需要支付一定的费用给中介 Agent M，即 $c=1$。测试的实验数据集是买方 Agent 和卖方 Agent 各有 4 组数据，4 组数据分别为 $D1$，$D2$，$D3$，$D4$。在表 1 中可以看到的是买方一栏中的保留价格 RP 参数没有进行初始化，是因为买方的期望效用函数中只与出售价格 p_0 有关，与保留价格 RP 没有关系。其次，是否接受一个协商议题与保留效用有关，与保留价格没有关系，因此，在初始化一些参数时，对保留价格初始化没有任何意义。

4.1　基于强化学习双边优化协商的实验结果与分析

在强化学习模型过程中，有时间信念函数、时间贴现率等重要的参数，大部分的学者主要研究协商的模型，当然也有孙天昊等人对参数进行研究，然而他们关注的是单议题价格，本文介绍价格和数量的相关性。进行仿真实验，观察强化学习中参数的不同取值对价格和数量的影响，结果验证了算法的可行性。选取最优的参数进行提议，以更好地符合现实生活中人的协调心理。

表 1 给出了 4 组实验数据(介绍 4 组数据是方便4.2 节使用)，这里采用第 3 组数据进行研究，其他数据的效果一样。

(1)买卖双方采用的时间信念是 t/T

图 2 分别示出卖方的协商曲线和买方的协商曲线。开始时买方的价格是 60，卖方的价格是 97。随着协商的进行，在第一次进行报价时，使用强化学习中的Q-learning算法进行报价。双方的第一次报价买方变成了64.7，卖方变成了 74.5。由于双方采用的时间信念函数 $sfb(t)$，$bfs(t)$ 都是 t/T，随着时间的增大，时间信念不断增加，由式(5)和式(6)可知平均回报值不断增加(双方对于协商的达成期望的要求越来越高)。由式(9)和式(10)可知，买方的出价不断降低，卖方的出价不断增加，从图中也可以看出直线的走势，同时，在第一次报价时就决定了是否能够达成协商。结论：在进行协商的过程中，双方的时间信念函数都采用这种函数，双方都是一种很强硬的态度，都不会进行妥协，显然，这样的结果是协商失败。

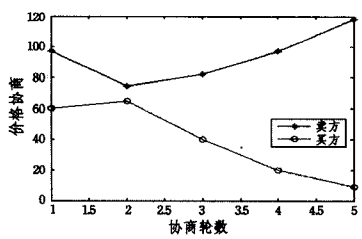

图 2　时间信念函数为增函数(t/T)

(2)双方采用的时间信念函数是 $1-t/T$

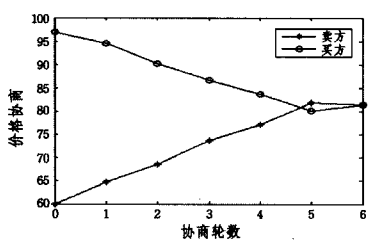

图 3　时间信念函数为减函数($1-t/T$)

从图 3 中能够清晰地看出双方是有交点的，能够达成协商。刚开始买卖双方都采用初始值，即文中给出的数据，买方的价格是 60，卖方的价格是 97。随着协商的进行，在第一次进行报价时，使用的是强化学习中的 Q-learning 进行报价。因此第一次的报价买方变成了 64.7，卖方变成了 94.5。由于双方采用的时间信念函数 $sfb(t)$，$bfs(t)$ 都是 $1-t/T$，随着时间的增大，时间信念不断减小，由式(5)和式(6)可知平均回报值不断减小(双方对于协商的达成期望的要求越来越低)。由式(9)和式(10)可知，买方的出价不断增加，卖方的出价不断降低，到达某个轮次，图中是在第 5 次出现买方的价格大于卖方，存在交易机会，后续的任务由中介 Agent 计算最终的交易价格。实验结果表明：采用时间信念函数为 $1-t/T$，在协商的过程中双方都进行让步，加快了协商的速度。

(3)常时间信念函数，分别采用的常时间信念为 0.3，0.5，0.7，0.9

在图 4 中，$s1$，$s2$，$s3$，$s4$ 分别表示常时间信念函数为 0.3，

0.5,0.7,0.9 卖方的协商曲线。从这个图中能够看出卖方初始的价格是 97,随着协商的进行,在第一次进行报价时,使用的是强化学习中的 Q-learning 算法。由于卖方采用的时间信念函数是常时间信念函数,从图中可以看出,卖方对于协商的达成期望的要求越来越高,以至于使得卖方的出价不断增加,某一方到达截止时间协商即失败。实验结果表明:采用时间信念函数为常时间函数,在协商的过程中卖家对达成期望的要求越来越高,这种报价的策略和时间信念函数为 t/T 的策略是比较相似的,只是时间信念函数为 t/T 的平均回报值随着时间的推进比常时间信念函数增加得要快。

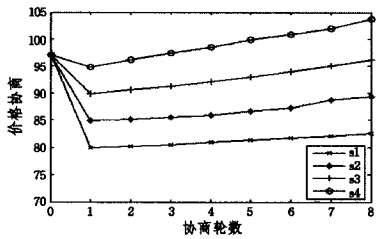

图 4 时间信念为常函数曲线图

因此,从上述的分析中能够看出,如果想要达成协商,那么双方就要进行妥协,也就是期望值不断地减小,每次买方的出价不断升高,卖方的出价不断降低,这样在截止时间之前是有可能达成协商的,这也比较符合现实生活中的协商过程。所以在协商模型中,采用时间信念函数 $1-t/T$ 进行研究,并将其运用到协商模型中,具有现实和可行性。

(4)对时间折扣率进行探讨

使用的数据是第 $D3$ 组数据,对卖方的时间折扣率分别为 0.9,0.7,0.5 进行研究,卖方的时间信念函数采用的是减函数,即 $1-t/T$。

在图 5 中,$s1$,$s2$,$s3$ 分别是时间折扣率 γ 值取 0.9,0.7,0.5 的协商曲线,从图形走势中能够看出,当时间折扣率 γ 为 0.5 时,卖方 Agent 的期望效用是比较小的。由式(5)可知,期望效用值(平均回报值)是关于时间折扣率 γ 的增函数。所以,当 γ 值取 0.5 时,卖方的价格下降得最快,即达成协商的愿望最强烈。一般的情况下,为了避免 Agent 妥协过快,损失更多自己的利益,时间折扣率 γ 取值为 0.9。

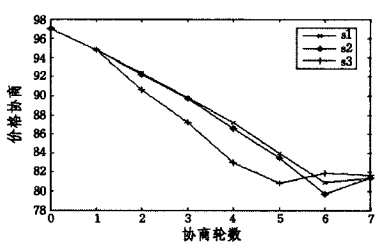

图 5 γ 取不同值时的曲线图

实验结果显示,在协商模型中,符合现实生活中人的报价心理,大部分的模型采用的时间信念函数为 $1-t/T$。为了避免 Agent 妥协过快,损失更多自己的利益,时间折扣率 γ 取值为 0.9。

4.2 基于中介 Agent 自适应学习能力的对手分类算法

根据强化学习双边优化协商的实验结果,采用时间信念函数为 $1-t/T$ 能够充分体现出人报价的心理状态,为了避免 Agent 妥协过快,损失更多自己的利益,时间折扣率 γ 取值为 0.9。因此本节实验选取时间信念函数为减函数、时间折扣率为 0.9 的强化学习双边优化协商算法进行改进。为了证明算法的有效性和可行性,本文选用的是强化学习双边优化协商算法(时间信念为减函数,折扣率为 0.9)和基于中介 A-gent 自适应学习能力的对手分类算法进行比较。

首先,对数据的协商过程进行详细分析,图 6 示出协商价格对于两个协商框架下的协商过程。由前面的分析可知,卖方 Agent 每次报价都会保证协商的数量是最大的,总会有 $q_s<q_b$,那么判断是否存在达成协商的标准就是变量价格 p。图 6 显示的就是随着协商轮次 T 不断地变化,关于价格的协商过程。

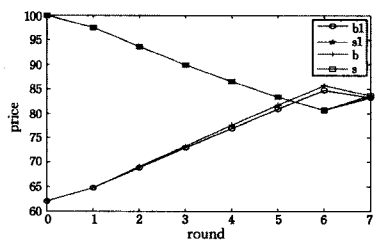

图 6 协商过程

本节实验采用的对比模型是基于强化学习双边优化协商模型。为了使研究好的协商模型更符合现实生活,采用买卖双方的期望还原率 γ 为 $\beta(t)\lambda_{max}$。将基于强化学习双边优化协商模型用符号 $\beta(t)\lambda_{max}$ 进行表示,同时,本节提出的模型基于中介 Agent 自适应学习能力的对手分类算法用 $bPNO$ 符号进行表示。

如果判断是否达成协商的标准是价格,那么这里对模型好坏进行评判的标准是期望效用值(利润值)。

在图 6 中,b 和 s 对应的是基于强化学习双边优化协商模型即 $\beta(t)\lambda_{max}$ 的买方和卖方。$b1$ 和 $s1$ 分别是中介 Agent 自适应学习能力的对手分类算法即 $bPNO$ 的买方和卖方。图的走势显示出了这两种协商模型对第二组数据 $D2$ 的协商过程:采用 $\beta(t)\lambda_{max}$ 和 $bPNO$ 两个协商模型,图中示意出了买方和卖方的协商议题价格 p 随着协商轮次 t 变化的协商曲线过程。分析图形可以看出,$\beta(t)\lambda_{max}$ 的协商模型中 $t=6$ 时,买方的价格高于卖方的价格,第一次显示出了交易机会;当 $t=7$ 时协商成功,总的协商次数为 8,最终的协商结果为〈89.7018,121〉,把协商的结果代入到买方的期望效用函数式(2)得到买方的期望效用值为 1750.58;而 $bPNO$ 的协商模型中,当 $t=6$ 时,买方的价格大于卖方的价格,第一次显示出了交易机会;当 $t=7$ 时协商成功,协商次数为 8,最终的协商结果为〈83.2074,122〉,把协商的结果代入到买方的期望效用函数式(2)得到买方的期望效用为 1805.69,即加入对手分类的算法,提高了期望效用值,买方获利更多。可见,本文所提的模型提高了买方的期望效用值,实验证明了该模型的有效性。

利用表 1 介绍的四组数据计算出在这两种协商模型价格的协商结果,再代入到买方的期望效用函数式(2)得出买方的期望效用值,对比如表 2 所列。分析这 4 组数据的期望效用值的协商结果可以得出:本文所提出的模型基于中介Agent自适应学习能力的对手分类算法即 $bPNO$,在买方效用值的方面都有所提高。

表 2 协商结果

N	Result		Utility	
	$\beta(t)\lambda_{max}$	bPNO	$\beta(t)\lambda_{max}$	bPNO
N1	〈89.39,114〉	〈88.986,114〉	1110.71	1156.6
N2	〈89.7018,121〉	〈83.2074,122〉	1750.58	1805.69
N3	〈80.9893,129〉	〈81.24,123〉	2030.89	2041.97
N4	〈72.65,140〉	〈72.224,141〉	3027.53	3074.81

图 7、图 8 分别代表的是随着截止时间的不断变化在这

两个协商模型下的协商次数、买方的期望效用的图形走势，这里选取了 10 个买方 Agent 的截止时间都为({12,15,17,20,22,25,27,30,32,35})。图 7 中，$t1$ 为选用 *bPNO* 模型时随着截止时间的增加买方协商轮次的变化曲线；t 为选用 $\beta(t)\lambda_{max}$ 模型时随着截止时间的增加买方协商轮次的变化曲线。能够看到 $\beta(t)\lambda_{max}$ 和 *bPNO* 这两个协商模型的协商次数随截止时间的增加不断增大，两者并没有太大的差距。在图 8 中，$u1$ 为选用 *bPNO* 模型时随着截止时间的增加买方的期望效用值的变化曲线；u 为选用 $\beta(t)\lambda_{max}$ 模型时随着截止时间的增加买方期望效用值的变化曲线。能够看出，$\beta(t)\lambda_{max}$ 和 *bPNO* 买方的期望效用值都是有波动的，但是也能够看出在本节所提基于中介 Agent 自适应学习能力的对手分类算法即 *bPNO* 中，买方的期望效用值都大于 $\beta(t)\lambda_{max}$ 模型中的期望效用值，那么，能够得出本节所提出的协商模型提高了买方的期望效用值。

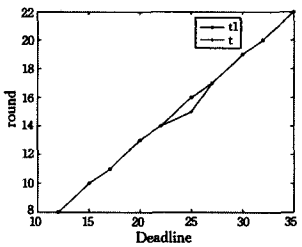

图 7　协商次数

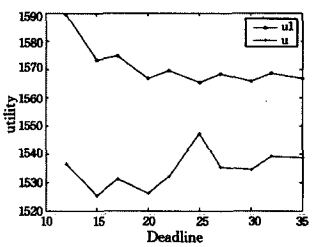

图 8　期望效用

4.3　实验小结

仿真实验结果表明，与性能较好的基于强化学习双边优化协商模型相比，基于中介 Agent 自适应学习能力的对手分类算法提高了一方的期望效用值，即对获取对手信息的那一方是有利的。本文只是关于议题的价格和数量，今后还会进一步对其进行改善，基于多线程的思想研究多议题并行协商的模型。

结束语　在电子商务活动中，协商是非常重要的组成部分。在人工智能方面，Agent 技术与多 Agent 技术的不断发展与成熟，给电子商务的智能化与自动化提供了技术支持，采用多 Agent 与 Agent 的自动协商技术为电子商务的发展做出了贡献。自动协商技术是计算机领域的研究热点，也取得了一定的研究硕果。本文针对协商过程中一些重要的参数和 Agent 自适应学习能力进行了改进和研究。

主要研究成果如下：

(1)针对采用增强学习的协商算法在协商过程中所有的信念函数是提前设置好的固定函数，本文对协商中的重要参数进行研究，主要是对时间信念函数的增函数、减函数、常时间函数以及时间折扣率等进行了讨论。实验结果表明：在强化学习双边优化协商模型中，模型采用的时间信念函数为 $1-t/T$，为了避免 Agent 妥协过快，损失更多自己的利益，时间折扣率 γ 取值为 0.9，这更符合现实生活中人的协商心理。

(2)针对已有的协商策略中介 Agent 没有自适应学习机制的缺点，在基于强化学习的双边优化协商模型上进行改进，使中介 Agent 存储协商者的历史信息并且评估协商的结果，买方从中介 Agent 获得卖方的评估协商结果，使得买方在动态变化的环境中采用卖方(对方)信念知识进行推理，动态地改变买方对卖方的信念函数，优化协商的报价策略。最后与第 4 节的性能较好的强化学习双边优化协商模型进行对比，实验结果表明，该算法使中介 Agent 具有自适应学习能力，同时使买方自主地提高协商能力，更高效地与对手进行协商，使自身的利益最大化。

本文的研究成果在一些方面改进了 Agent 的自动协商的模型，但是还是存在着一些地方需要完善，如本文中研究的是价格和数量的相关性，未来是可以增加更多的协商议题，基于多线程的思想研究多议题并行协商的模型。

参 考 文 献

[1] BAARLAG T, HENDRIKX M J C, HINDRIKS K V, et al. Learning about the opponent in automated bilateral negotiation: a comprehensive survey of opponent modeling techniques[J]. Autonomous Agents and Multi-Agent Systems, 2015, 30(5): 849-898.

[2] ILANY L, GAL Y. Algorithm selection in bilateral negotiation [J]. Autonomous Agents and Multi-Agent Systems, 2016, 30 (4): 697-723.

[3] ZHANG X S, KLEIN M, MARSAMAESTRE I. Scalable Complex Contract Negotiation with Structured Search and Agenda Management[C]// AAAI Conference on Artificial Intelligence. 2014.

[4] ZHANG Hua-xiang, HUANG Shang-teng. Agent Negotiation Model Based on Reinforcement learning[J]. Computer Engineering, 2004, 30(10): 137-139. (in Chinese)
张化祥，黄上腾. 基于增强学习的代理谈判模型[J]. 计算机工程，2004，30(10)：137-139.

[5] SUN T H, CHEN F, ZHU Q S. Reinforcement learning negotiation strategy based on Bayesian classification[J]. Chinese Journal of Computer Science, 2011, 38(9): 227-229. (in Chinese)
孙天昊，陈飞，朱庆生. 基于贝叶斯分类的增强学习协商策略[J]. 计算机科学，2011，38(9)：227-229.

[6] ZHANG Lin-lan, SONG Hai-gang, CHEN Xue-guang, et al. A simultaneous multi-issue negotiation through autonomous agents[J]. European Journal of Operational Research, 2010, 210 (1): 95-105.

[7] SUN Tian-hao, DENG Jun-kun, CAO Feng, et al. Reinforcement Learning Negotiation Strategy based on Opponent Classification [C]// The International Conference on Computer Science and Service System (CSSS 2011). Nanjing, China, 2011: 3987-3989.

[8] SUI Xin, CAI Guo-yong, SHI Lei. Multi-agent negotiation strategy and algorithm based on Q-Learning[J]. Chinese Journal of computer engineering, 2010, 36(17): 198-200. (in Chinese)
隋新，蔡国永，史磊. 基于 Q-强化学习的多 Agent 协商策略及算法[J]. 计算机工程，2010，36(17)：198-200.

[9] CHEN Li-hong, DONG Hong-bin, HAN Qi-long, et al. Bilateral Multi-issue Parallel Negotiation Model Based on Reinforcement Learning[C]// The 14th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2013, LNCS 8206). 2013.

[10] DIAMAH A, MOHAMMADIAN M, BALACHANDRAN B. Fuzzy utility and inference system for bilateral negotiation[C]// 2012 International Conference on Uncertainty Reasoning and Knowledge Engineering. 2012: 115-118.

[11] CHEN Li-hong, DONG Hong-bin, ZHOU Yang. A reinforcement learning optimized negotiation method based on mediator agent[J]. Expert Systems with Applications, 2014 (41): 7630-7640.

[12] 普尔. 人工智能：计算 Agent 基础[M]. 董红斌，等译，北京：机械工业出版社，2015.