基于改进增强学习算法的双边多协议协商策略

张 科 罗 军 邓俊昆
(重庆大学计算机学院 重庆 400044)

摘 要 针对传统增强学习算法存在妥协过快导致自身效用降低的缺点,通过设计改进增强学习算法的双边多议题协商模型,引入期望还原率,还原Agent的期望,从而提高协商解的质量。通过实验分析了期望还原率不同取值对协商的影响,并对传统增强学习协商策略、基于时间的协商策略和改进增强学习协商策略的协商效果做了对比。实验表明,在协商次数允许的范围之内,基于期望还原率的改进增强学习算法在双边多议题协商中能够提升双方的效用。

关键词 协商策略,增强学习,期望还原率,双边多议题
中图法分类号 TP301.6 **文献标识码** A

Bilateral Multi-protocol Negotiation Strategies Based on Reinforcement Learning

ZHANG Ke LUO Jun DENG Jun-kun
(College of Computer, Chongqing University, Chongqing 400044, China)

Abstract Traditional reinforcement learning negotiation strategy has the shortcoming of compromising too fast and reduces the utility of agent. Aiming at this problem, improved reinforcement learning bilateral multi-issue negotiation strategy which imports expectation restoration rate to restore the expectation of agent can improve the quality of the negotiation result. This paper analysed the influence of different expectation reduction rate on negotiation and contrasted traditional reinforcement learning negotiation strategies, time-based negotiation strategy and the proposed enhance learning negotiation strategy consultation. The result shows that negotiation strategy can get higher bilateral utility within allowing negotiation turns.

Keywords Negotiation strategy, Reinforcement learning, Expectation restoration rate, Bilateral multi-issue negotiation

将学习机制引入基于Agent的电子商务协商中,也就是在协商过程中学习对手的信念、偏好以及协商环境知识,使得Agent适应动态变化的环境,通过不断调整自身信念,采取不同策略与对手进行协商,使利益最大化[1]。

目前已有多种机器学习的方法应用到协商当中,常见的有贝叶斯学习、遗传算法、增强学习、支持向量机等等。贝叶斯学习主要是通过不断更新对周边环境的信念和其他agent的信息,采取有利于自己的方式协商[2]。文献[3]通过构造一个带约束的最优化问题,将混合遗传算法应用到双边多议题协商中,提高了协商解的质量。文献[4]采用支持向量机训练对手的态度,针对不同对手偏好的不同采用不同的协商策略。

增强学习通过学习最优动作获得最好的回报,增强学习的基本思想是对所希望的结果给予奖励,对不希望的结果给予惩罚,逐渐形成一种趋向于好的结果的条件反射[5]。文献[6]将增强学习中的Q学习算法应用到自动协商中,提出了基于时间信念、价格信念的协商模型,但是在其所提模型中,协商双方妥协较快,不符合现实生活的场景。文献[7]将贝叶斯学习与增强学习算法结合起来,通过贝叶斯学习进行更新信念,然后用增强学习算法对协商进行求解。

文献[8]针对传统增强学习的妥协过快的缺点,提出了期望还原率加以改进,并用实验证明了其有效性,但是其所做的改进是针对单议题的,有一定局限性。本文设计一个双边多议题的协商模型,将基于期望还原率的增强学习算法应用到该模型中,通过与传统增强学习协商策略、基于时间的协商策略进行对比,验证了基于期望还原率的增强学习协商策略的有效性。

1 双边多议题协商框架

一般自动协商模型具备如下特性:(1)能够有效描述问题;(2)能够支持学习;(3)能够为Agent提供灵活的协商提议;(4)在资源有限的情况下完成协商。基于以上特性,定义协商框架(framework of negotiation)如下:

$FN=\langle A,T,X,D,W,U\rangle$

其中:

(1)A代表协商Agent的集合。$A=\{A_1,A_2\}$,因为本文讨论的是双边多议题协商,所以只有两个agent,也可看作$\langle buyer,seller\rangle$。

(2)T是指协商轮次,代表协商的时间期限,T_b和T_s分别代表买卖双方的最大协商次数。

(3)X是协商的议题集合。$X=\{X_1,X_2,\cdots,X_n\}$,议题

到稿日期:2013-03-16 返修日期:2013-06-23 本文受中央高校基本科研业务费科研专项项目(CDJZR10180014)资助。

张 科(1987—),男,硕士生,主要研究方向为机器学习、语义网,E-mail:oscarzhangke@gmail.com(通信作者);罗 军(1961—),男,副教授,硕士生导师,主要研究方向为数据库及其办公系统自动化、语义网与知识管理系统;邓俊昆(1988—),男,硕士生,主要研究方向为机器学习、自动协商。

X_i 就是所要协商的对象，比如价格、数量、地点等等。

(4)D 是议题的取值范围。$D=\{D_1,D_2,\cdots,D_n\}$，D_i 指某个议题的取值区间。

(5)W 是议题的权重。$W=\{W_1,W_2,\cdots,W_n\}$，其中，$\sum_{i=1}^{n} W_i=1$，不同权重代表不同议题对 Agent 的重要程度。

(6)U 是 Agent 的效用。$U_i=\sum_{j=1}^{N} w_j^i v_j^i(x_j^i)$，$v_j^i$ 是评分函数，或称子效用函数。其中当 v_j^i 单调递减时，$v_j^i=\frac{high_j^i-x_j^i}{high_j^i-low_j^i}$，当 v_j^i 单调递增时，$v_j^i=\frac{x_j^i-low_j^i}{high_j^i-low_j^i}$。

协商结果评判标准：

对一个协商模型的评价主要包括以下 3 个方面：协商效用(包括己方效用、对方效用及效用和)、协商时间效率、建议生成时间[9]。这个 3 个方面相互联系而又相互制约。如果要达成较高的协商效用，一般需要深入探讨，双方可能需要长时间的协商，这时协商时间效率就会下降，如果用到较复杂的协商模型，则建议生成时间变长，而在另一方面协商时间可能会变短。一般来说，效用优先，在协商允许的时间内效用和越高越好。

2 3 种协商策略及协商算法

2.1 传统基于增强学习协商策略

传统增强学习算法是对单议题进行协商的，议题为价格，买卖双方的价格取值区间为$[P_b^{\min},P_b^{\max}]$和$[P_s^{\min},P_s^{\max}]$。时间信念是指 Agent 认为随着时间的推进，对方接受其报价的概率，买卖双方的时间信念分别记为${}^bb^s(t)$、${}^sb^b(t)$。价格信念是指 Agent 对成交价格在其报价区间内概率分布的认识。买卖双方的价格信念分别记为${}^bp^s$、${}^sp^b$。

Q 函数的基本定义如下：

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

式中，γ 为时间贴现率，$\delta()$ 为状态转移函数。

在自动协商中，Agent 每次报价可以看成是在某一个状态选择行动来实现状态的迁移。如在第 t 次报价时的状态为 $s(t)$，报价为 $p(t)$，报价后进入另一个状态 $s(t+1)$。

若 Agent 在第 t 次达成协商，则：

卖方 Agent 回报

$$Q_e^s=\int_{p_s^{\min}}^{p_s^{\max}}(p^T-p_s^{\min})^sp^bd_{p^T} \tag{2}$$

买方 Agent 回报

$$Q_e^b=\int_{p_b^{\min}}^{p_b^{\max}}(p_b^{\max}-p^T)^bp^sd_{p^T} \tag{3}$$

卖方 Agent 第 t 阶段 Q 值的平均期望为：

$$\overline{Q_s}(s(t),p(t))=(\sum_{i=t}^{T_s}{}^sb^b(i)\gamma^{i-t}Q_e^s)/(T_s-t+1) \tag{4}$$

买方 Agent 第 t 阶段 Q 值的平均期望为：

$$\overline{Q_b}(s(t),p(t))=(\sum_{i=t}^{T_b}{}^bb^s(i)\gamma^{i-t}Q_e^b)/(T_b-t+1) \tag{5}$$

最后，得到卖方 Agent 的报价策略为：

$$p_s(t)=p_s^{\min}+\overline{Q_s}(s(t),p(t)) \tag{6}$$

买方 Agent 的报价策略为：

$$p_b(t)=p_b^{\max}-\overline{Q_b}(s(t),p(t)) \tag{7}$$

2.2 基于期望还原率的增强学习协商策略

文献[8]指出，由于过多考虑未来期望，传统增强学习算法有着妥协过快的缺点，其改进如下：

定义 1 期望还原率 α 表示 Agent 对原始期望的还原程度。α_s、α_b 分别对应卖家和买家的期望还原率。

引入 α 之后，对式(6)、式(7)进行修改。

$$p_s(t)=p_s^{\min}+\alpha_s\overline{Q_s}(s(t),p(t)) \tag{8}$$

$$p_b(t)=p_b^{\max}-\alpha_b\overline{Q_b}(s(t),p(t)) \tag{9}$$

α 主要有两种取值方式：

$$\alpha=\begin{cases}\alpha_{\max}\\ \beta(t)\alpha_{\max}\end{cases}$$

其中，$\alpha_{\max}=(p^{\max}-p^{\min})/\overline{Q}(s(t),p(t))$，$\beta(t)=(1-t/T)\in[0,1]$。

2.3 基于时间的协商策略

在协商中，时间是一个很重要的因素，也是协商双方态度的一个体现。作为理性的协商 Agent，在协商刚开始时应该使自己的利益最大化，随着协商的不断推进，Agent 才不断地降低自己的期望。在基于时间的协商中，Agent 期望效用是随时间增加而递减的一个函数，其提议决策模型如下：

$$x_{a\to b}^t[j]=\begin{cases}x_{a\to b}^{t-1}-f_j^a(t)(high_j^a-low_j^a), & v_j^a \text{ 是严格递增函数}\\ x_{a\to b}^{t-1}+f_j^a(t)(high_j^a-low_j^a), & v_j^a \text{ 是严格递减函数}\end{cases}$$

$x_{a\to b}^t[j]$代表 Agent a 在第 t 轮协商时对第 j 个议题的提议，其中：

$$x_{a\to b}^1[j]=\begin{cases}high_j^a, & \text{当 } v_j^a \text{ 是严格递增函数时}\\ low_j^a, & \text{当 } v_j^a \text{ 是严格递减函数时}\end{cases}$$

$f_j^a(t)$是随时间变化的一个函数，为了一般性讨论，本文采用的效用函数随时间均匀递减。$f_j^a(t)=1/t_{\max}^a$。

2.4 协商算法步骤

本文采用交替提议方式，轮流报价，算法顺序执行。

1. 选定要进行协商的 Agent，确定协商议题，初始化协商次数、议题取值区间、议题权重等等；

2. 由卖方向买方发送第一个提议，买方收到提议后，首先判断是否到达最大时限，如果是，协商失败。

3. 买方对卖方的提议进行判断，检查约束条件，如果所有的议题提议都落在可以接受的区间内，且效用不小于买方下次提议的效用，就接受该提议，协商成功。否则，根据协商模型提供的协商策略提出反提议。

4. 卖方接到买方提议时，卖方首先对己方时限进行判断，若已达最大时限，协商失败。否则，检查约束条件，如果对所有的议题提议都落在可以接受的区间内，且效用不小于卖方下次提议的效用，就接受该提议，协商成功。否则，根据协商模型提供的协商策略提出反提议，转 3。

3 实验

为了验证协商策略的有效性，本文对传统增强学习协商策略、基于时间的协商策略和基于期望还原率的增强学习策略进行对比研究。只有当协商双方对不同议题有着不同偏好时，才能使得协商双方对多个议题进行协商时实现共赢，本文就是基于这个角度设计实验数据的。

(1)随机抽取一组实验数据，见表 1，其中 $\gamma=0.9$，${}^bp^s=1/(p_b^{\max}-p_b^{\min})$，${}^sp^b=1/(p_s^{\max}-p_s^{\min})$，${}^bb^s(t)=(1-t/T_b)$，${}^sb^b(t)=(1-t/T_s)$，$\alpha=(1-t/T)\alpha_{\max}$。在这组数据中，买卖双方对待不同的议题有着不同偏好，权重刚好相反，即当 Agent 对某个不重要的议题做出较大让步时，会给对方带来较高的效

用，而自身的效用不会损失太多。

表1 实验数据

议题	T_b	T_s	D_b	D_s	W_b	W_s
X1	20	20	[20,10]	[50,15]	0.7	0.4
X2	20	20	[40,25]	[30,20]	0.3	0.6

实验结果效果图如图1所增。

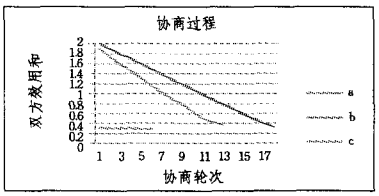

图1 协商过程

图1中，a、b、c 3条曲线分别对应传统增强学习协商策略、基于时间协商策略、基于期望还原率的增强学习协商策略的协商双方效用和曲线。

实验结果表明，基于期望还原率的增强学习算法有着更高的效用：

a)从图1可以看出，与传统增强学习协商策略相比，基于期望还原率的增强学习协商策略效用和更高。传统增强学习算法虽然只需要6次就可以达成协商，但效用和较低。

b)与基于时间的协商策略相比，基于期望还原率的协商策略效用和略高，同时协商时间也更短，只需要13次，而基于时间的协商需要17次。

在第1节中已经提到对协商结果进行评判时优先考虑效用和，事实上，当协商次数在允许的范围内时，效用越大，模型可用性越高；另外，图1中协商成功时效用和都比较低，这是因为各个议题的议价区间差距较大，双方都能接受的议题值有限，所以在达成协商时效用和也不高。

(2) 实验数据采用表1中的数据，当期望还原率取不同的值时，实验效果如图2所示。

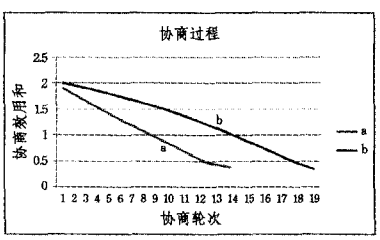

图2 不同期望还原率协商对比图

图2中，a、b分别对应期望还原率为$(1-t/T)\alpha_{max}$、α_{max}时的情形。从图2中可以看出，当$\alpha=(1-t/T)\alpha_{max}$时，协商的速度更快，只需要13次就能达成协商，协商效用和更高(0.39)，而$\alpha=\alpha_{max}$时，协商要在第19次才能达成，达成时的效用和为0.36。这是因为当α取α_{max}时，算法对期望的还原程度较大，但是若买卖双方都采用此策略，就会对协商造成拖延，虽然协商效用和也较高，但在时间上降低了协商效率。而α取$(1-t/T)\alpha_{max}$时，Agent既能利用增强学习算法协商较快的优点，又避免了过度妥协，在时间和效用上都取得了不错的效果。

(3)本文进行了1000次实验，并对比了实验结果，实验数据自动随机生成，期望还原率取$(1-t/T)\alpha_{max}$。实验数据的区间如表2所列。

表2 实验数据

名称	区间	名称	最小值区间	最大值区间
T_s	[20,22]	卖家议题X1	[16,18]	[42,45]
T_b	[20,22]	买家议题X1	[8,12]	[23,25]
W_{s1}	[0.1,0.9]	卖家议题X2	[18,20]	[28,32]
W_{b1}	[0.1,0.9]	买家议题X2	[23,27]	[38,42]
W_{s2}	[0.1,0.9]	W_{b2}	[0.1,0.9]	

协商结果如表3所列。

表3 协商结果

算法名称	协商成功次数	效用和较高次数
传统增强学习算法	850	252
基于时间算法	920	198
改进增强学习算法	950	550

其中，协商和较高次数是指对1000次协商中3种协商策略所得效用和较高策略的统计。从表3中可以看出，改进增强学习算法有95%的协商成功率，并且在55%的情况下协商和更高。这也在一定程度上说明了改进算法的优越性。

总体来说，传统增强学习算法协商速度快，但是效用和较低，由于未充分利用协商次数，导致在议价区间差别不大时不能进行充分的协商，体现为协商不公平。在议价区间差别较大时，会造成双方效用和较低。而改进后的增强学习算法在进行双边多议题协商时能够考虑到协商次数、协商态度这些问题，从而提高了协商成功率，提升了协商解的质量。

结束语 本文将基于期望还原率的增强学习算法应用到双边多议题的协商中，对比了传统增强学习算法和基于时间的协商策略，从理论和实验上证明了所提策略的有效性。该策略继承了增强学习协商较快的优点，同时又避免了过度妥协的弊端，能够使协商结果在时间和效用两方面都有所提升。下一步的研究方向是结合对手分类和增强学习策略的双边多议题协商。

参考文献

[1] Park S, Yang Sung-Bong. An efficient multilateral negotiation system for pervasive computing environments [J]. Engineering Applications of Artificial Intelligence, 2008, 21: 633-643

[2] Zeng D, Sycara K. Bayesian learning in negotiation[J]. Int'l J. Human-Computer Studies, 1998, 48: 125-141

[3] 李剑，牛少彰. 一种基于混合遗传算法的双边多议题协商[J]. 北京邮电大学学报，2009，32(2)：1-4

[4] 程昱，高济，古华茂，等. 基于对手态度学习的协商决策模型[J]. 浙江大学学报：工学版，2008，42(10)：1676-1

[5] Mitchell T M. 机器学习[M]. 曾华军，等译. 北京：机械工业出版社，2009

[6] 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

[7] Chen Pei-you, Li Yi-jun, Li Xing. The research on E-business-oriented Automatic Negotiation System based on faithful and dynamic Q-study[C]// Chinese Control and Decision Conference. Shenyang, 2008

[8] 孙天昊，邓俊昆，陈飞，等. 基于增强学习协商策略的研究及优化[J]. 计算机工程与应用，2012，48(23)：44-46

[9] 艾解清. 双边多议题自动协商研究[D]. 杭州：浙江大学，2011：46-47

[10] 罗志伟. 协同设计系统中图形协同与网络协商的实现[J]. 重庆理工大学学报：自然科学版，2012，26(7)：27-33