基于神经网络的 Agent 增强学习模型*)

唐亮贵[1,2] 刘 波[1] 唐 灿[1] 程代杰[2]

(重庆工商大学计算机学院 重庆 400067)[1] (重庆大学计算机学院 重庆 400030)[2]

摘 要 在深入分析 Agent 决策过程中状态与行为空间的迁移与构造的基础上,设计了 Agent 基于强化学习的最优行为选择策略和 Agent 强化学习的神经网络模型与算法,并对算法的收敛性进行了证明。通过对多 Agent 电子商务系统中 Agent 竞价行为的预测仿真实验,验证了基于神经网络的 Agent 强化学习算法具有良好的性能和行为逼近能力。

关键词 Agent,强化学习,神经网络,Markov 决策过程

An Agent Reinforcement Learning Model Based on Neural Networks

TANG Liang-Gui[1,2] LIU Bo[1] TANG Can[1] CHENG Dai-Jie[2]

(College of Computer Science, Chongqing Technology and Business University, Chongqing 400067)[1]

(College of Computer Science and Technology, Chongqing University, Chongqing 400044)[2]

Abstract This paper thoroughly analyzes the transfer and construction of the state-action space of the agent decision-making process, discusses the optimal strategy of agent's action selection based on Markov decision-making process, and designed a neural networks model for the agent reinforcement learning, and designed the agent reinforcement learning based on neural networks. By the simulation experiment of agent's bid price in Multi-Agent Electronic Commerce System, validated the Agent Reinforcement Learning Algorithm Based on Neural Networks has very good performance and the action impending ability.

Keywords Agent, Reinforcement learning, Neural networks, Markov decision-making process

1 引言

强化学习(Reinforcement Learning)是一类通过试错并与环境交互获得反馈,求解序贯优化决策问题的机器学习方法,其主要特征是利用不确定的环境奖赏值来发现最优行为策略。近年来,作为人工智能和机器学习领域的研究热点之一,强化学习在理论和算法上已取得了大量的研究和应用成果[1~3]。

目前已提出的强化学习算法如 TD(λ)[4]、Q-学习[5]和 Sarsa-学习[6]算法等,基本上都是以 Markov 决策过程(MDP)和基于动态规划的值函数迭代计算为基础,并通过多项式基函数、决策树或多层前馈神经网络等值函数逼近器来解决强化学习中大规模连续状态空间计算的"维数灾难"(Curse of Dimensionality)问题,实现强化学习的泛化(Generalization)。

对于基于神经网络的强化学习的研究,主要集中在以下两个方面[3,10,11]:一是以强化学习系统为框架,将强化学习看作一个控制系统,通过神经网络对强化学习系统中输入等参量进行处理,优化强化学习行为;二是将强化学习算法中的 Q 值迭代思想引入神经网络的联接权值调整与修正中,特别是输出层权值的学习与优化,从而提高神经网络的逼近能力。

在多主体系统(Multi-Agent Systems, MAS)中,由于信息的分布性,单个行为主体(Agent)能够获取的关于其他主体的信息很少。甚至在主体获得其他主体以前的行为信息的情况下,其他主体的行为还可能因为他们的学习而随时改变当前的行为策略。因此,在问题求解过程中,单个主体不断强化自身能力和最大化收益的同时,还应协同其他行为主体形成一种联合求解的智能群体,即在竞争与合作的过程中要根据所处的状态、所求解的任务以及环境的反馈与奖赏等信息,去预测其他主体的行为策略以及联合状态等,以提高 Agent 强化学习的效率和行为逼近能力。

2 基于 MDP 的 Agent 行为选择策略

定义 1 行为 Agent 具有连续状态空间和离散行为空间的 Markov 决策过程(MDP)定义为一个四元组$\langle S,A,R,T\rangle$,其中 S 为连续状态空间,A 为有限的离散行为空间,$r\in R: S\times A\to R$ 为奖赏函数,$p\in P: S\times A\to P$ 为状态转移函数,P 是状态空间上的概率分布集合。通常在行为选择策略 π 下,概率分布 P 可能随时间发生变化,即行为策略是非平稳的。

定义 2 (状态值函数) 设 Agent 的行为选择策略为 π,系统状态为 s_t,$r_t\in R$ 是在状态 s_t 下获得的奖赏值,$0\leqslant\beta\leqslant 1$ 是折扣因子,则其状态值函数 $v(s_t,\pi)$ 定义为:$v(s_t,\pi)=\sum_{t=0}^{\infty}\beta^t E(r_t/\pi,s_t)$,如果在给定行为选择策略 π 下的一种动作 $a_t\in A$,使初始状态 s_t 转移到状态空间 s_{t+1},则上式可改写成:

$$v(s_t,\pi)=r(\pi(s_t))+\beta\sum_{s_{t+1}\in S}p(s_t,a_t,s_{t+1})v^{\pi}(s_{t+1}) \quad (1)$$

动态规划理论保证至少存在一种优化行为策略 π^*,对 $\forall s\in S$,有

$$v(s_t,\pi^*)=\max_a\{r(\pi(s_t))+\beta\sum_{s_{t+1}}p(s_t,a_t,s_{t+1})v(s_t,\pi^*)\} \quad (2)$$

$v(s_t,\pi^*)$称为最优状态值函数。

定义 3 (行为值函数) 设 Agent 的行为选择策略为 π^*,$a_t\in A\subseteq\pi^*$,在 t 时刻系统状态为 s_t,在动作 a_t 下系统状态转移为 s_{t+1},r_t 是在动作 a_t 下获得的奖赏值,$0<\beta<1$ 是折

*)重庆市重点科技攻关资助项目(CSTC,2005AC2090)、重庆市自然科学基金资助项目(CSTC,2004BB2167;CSTC,2006BB2249)、重庆市教委科技项目(KJ060704)。**唐亮贵** 副教授,博士生,主要研究方向:智能多代理系统与分布式计算、智能电子商务理论与技术、网络计算与应用。

扣因子，则其行为值函数 $Q(s_{t+1},a_t)$ 定义为

$$Q(s_{t+1},a_t)=r_t+\beta\sum_{s_{t+1}\in S}P(s_t,a_t,s_{t+1})v(s_{t+1},\pi^*) \quad (3)$$

系统在状态 s_t 时执行动作 a_t，且此后按最优动作序列执行，则其折扣累计强化值为

$$Q(s_{t+1},a_t)=r_t+\beta\max_{a\in A}\{Q(s_{t+1},a)\} \quad (4)$$

由前面的定义知，只有在得到最优策略的前提下该等式才成立，在学习阶段等式不成立，其更新规则是基于瞬时差分(Temporal Difference Method，TD)方法按下式进行的：

$$Q(s_t,a_t)=\begin{cases}Q(s_{t-1},a_{t-1})+\eta_t[r_t+\beta\max_{a\in A}\{Q(s_{t-1},a)\}-\\ \quad Q(s_{t-1},a_{t-1})],s=s_t\in S,a=a_t\in A\\ Q(s_{t-1},a_{t-1}),\text{其他}\end{cases} \quad (5)$$

其中 s_t 是当前状态，a_t 是选定的动作，η_t 是学习因子。

根据上述分析，动作选择应以较大概率从行为策略集合中选择最大行为值函数的行为元素，故行为选择概率对值函数估计的变化不具有连续性，常采用统计物理学中的 Boltzman 分布的近似贪心且连续可微的行为选择策略。设在状态 s_t 下，系统的行为集合为 $A=\{a_t,i=1,2,\cdots,n\}$，$Q(s_t,a_t)$ 为行为值函数的估计，则行为选择概率为

$$p(s_t,a_t)=\frac{e^{Q(s_t,a_t)/T}}{\sum_{a\in A}e^{Q(s_t,a)/T}}=\frac{1}{\sum_{a\in A}e^{\frac{Q(s_t,a)-Q(s_t,a_t)}{T}}} \quad (6)$$

其中 $T>0$ 为温度参数或能量参数。

当 $T\to 0$ 时，上述行为选择策略近似于贪心策略。

3 基于神经网络的 Agent 强化学习

强化学习是学习怎样去做，即如何将状态向行为进行映射，以最大化其回报，通常定义在 Markov 决策过程优化控制模型的框架上。通过神经网络的值函数逼近器可以降低强化学习中大规模连续状态空间计算的维数，提高 Agent 的学习效率。

3.1 Agent 强化学习的神经网络模型与算法

不失一般性，设在给定的时间段内 Agent 基于 MDP 的行为集元素个数为有限数 N，则可采用 N 层前馈神经网络来设计 Agent 的强化学习模型。该神经网络模型由 N 个学习单元(Learning Unit，LU)组成，每个学习单元分别逼近每个行为值函数 $Q(s_i,a_i)$。设神经网络学习单元 LU_i 的输入由 MDP 的瞬时状态 s_i、行为 a_i、环境奖赏值 r_i 和 $Q_{i-1}(i=1,\cdots,n)$组成，输出是 MDP 行为值函数 $Q_i(s_i,a_i)$的值，则 Agent 强化学习的神经元网络模型可如图 1 所示方式设计，其中 LU_i 是第 $i(i=1,2,\cdots,n)$个学习单元，其中 r_0、s_0、a_0、Q_0 为初始值。

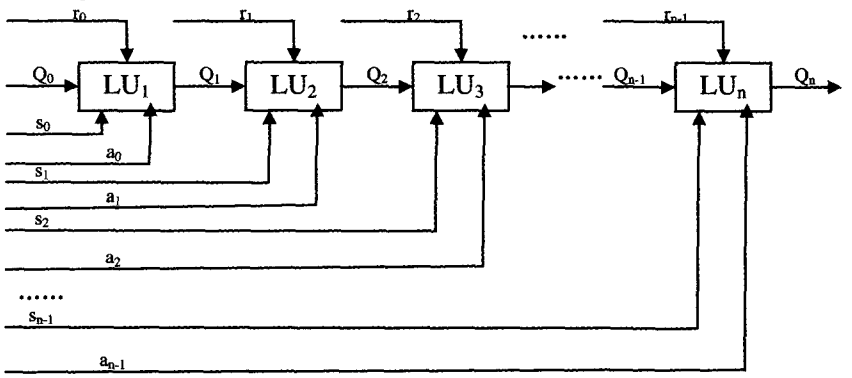

图 1 Agent 强化学习的神经元网络模型

设对每一个学习单元 $LU_i(i=1,2,\cdots,n)$所输入的瞬时状态 s_i 由状态向量 $X'=\{x_1,x_2,\cdots,x_m\}$构成，学习单元隐含层权值为 q_{ij} $(i=1,2,\cdots,m;j=1,2,\cdots,l)$，$m$ 为输入维数，l 为隐含层单元数，则输入向量为 $X=\{x_1,x_2,\cdots,x_m;r_i,a_i,Q_{i-1}\}$，输出向量为 y_i，输出层权值向量为 w_i。学习单元 LU_i 的 Q-网如图 2 所示。

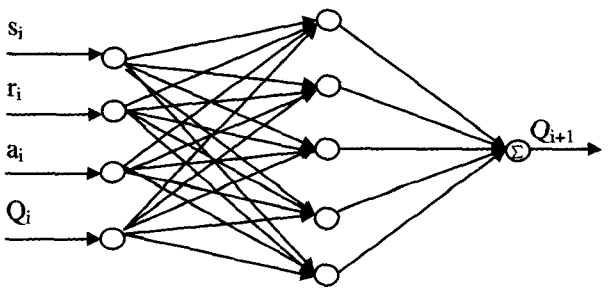

图 2 神经网络学习单元 LU_i 结构图

因此，Agent 的 MDP 行为值函数在时刻 i 的估计具有如下形式：

$$Q(s_i,a_i)=(w_i)^Ty_i=\sum_{j=1}^{n}w_i^jy_i^j \quad (7)$$

根据 Q 函数的定义知，只在最优策略下才能取得最优行为值 Q^*，在学习阶段 Q 值迭代存在误差。由瞬时差分(TD)法，设相应的 TD 误差为

$$\delta_t=r_t+\beta\max_{a\in A}\{Q(s_{t-1},a)\}-Q(s_{t-1},a_{t-1}) \quad (8)$$

为了使 Agent 行为最优，就应尽量使误差最小，因此可以通过 Bellman 残差平方和的最小化来实现。

设行为选择概率为 $p(s_i,a_i)$，$E[\cdot]$为定义在概率分布 P 上的数学期望，则 Bellman 残差平方和指标定义为

$$\tilde{E}=\frac{1}{2n}\sum_{s_i}\sum_{a_i}E[r(s_i,a_i)+\beta\sum_{j=i+1}p(s_j,a_j)Q(s_j,a_j)-Q(s_i,a_i)]^2 \quad (9)$$

用随机梯度下降法极小化 $\tilde{E}$，则有

$$\frac{\partial\tilde{E}_t}{\partial w_i}=p(s_{t+1},a_{t+1})\delta_t\left[p(s_{t+1},a_{t+1})\frac{\partial\delta_t}{\partial w_i}+\delta_t\frac{\partial p(s_{t+1},a_{t+1})}{\partial w_i}\right] \quad (10)$$

其中，

$$\frac{\partial\delta_t}{\partial w_t}=\beta\frac{\partial Q(s_{t+1},a_{t+1})}{\partial w_i}-\frac{\partial Q(s_t,a_t)}{\partial w_i} \quad (11)$$

$$\frac{\partial p(s_{t+1},a_{t+1})}{\partial w_i}=\frac{-1}{Tp^2(s_{t+1},a_{t+1})}\sum_{a\in A}[e^{(Q(s_{t+1},a)-Q(s_{t+1},a_{t+1}))/T}$$

$$\left(\frac{\partial Q(s_{t+1},a)}{\partial w_i}-\frac{\partial Q(s_{t+1},a_{t+1})}{\partial w_i}\right)] \tag{12}$$

$$\frac{\partial Q(s_{t+1},a)}{\partial w_i}=\begin{cases}y_i, \text{若 } a=a_i\\ 0, \text{若 } a\neq a_i\end{cases} \tag{13}$$

采用上述随机梯度下降的神经网络输出层权值迭代公式为

$$\Delta w_i=-\eta_t\frac{\partial \tilde{E}_t}{\partial w_i} \tag{14}$$

因此 Agent 基于神经网络的强化学习算法为：

算法 1 基于神经网络的 Agent 强化学习算法(ANNRL)

For all $LU_i, i\leqslant N$
  Initialize the joint power value w
  Initialize learning parameters s_0, r_0, Q_0;
  $t\leftarrow 0$;
  While $t<$MaxNum
    asuppose current state is s_t,
    Choose action a_t^i based on $p(s_t,a_t^i)$;
    Observe the state s_{t+1} and reward $r(s_t,a_t)$;
    Choose action a_{t+1}^i based on p;
    Update joint power value While $j\leqslant M$
      Computing $\frac{\partial \tilde{E}_t}{\partial w_i}$;
      Update every output layer joint power value by $\Delta w_i=-\eta_t\frac{\partial \tilde{E}_t}{\partial w_i}$;
      $j\leftarrow j+1$;
    End While
    Computing hiden layer joint power value by BP;
    $t\leftarrow t+1$;
  End While
  Computing Q_i
  $i\rightarrow i+1$;
End For.

其中，N 为学习单元数，MaxNum 是根据状态观测所确定的一个算法终止条件，M 为输出层连接数，$0<\eta_t<1$ 为学习因子。

根据算法 1，Agent 在实现对自己的 Q 值更新的同时，完成对信念的修正，从而降低 Agent 联合状态空间和联合行为空间的维度，并通过基于神经网络的强化学习，提高 Agent 在协同工作中的效率。

3.2 算法的收敛性

在一定的假设条件下，上述算法通过学习，可以收敛到优化的和均衡的 Q 值 $Q^*(s,a)$。

收敛定理 假设每个状态-行为对都被无限频繁地访问；学习因子 $\eta_n(s_i,a_i)$ 是有界的，如 $\eta_n(s_i,a_i)$ 满足下述条件：

$$\sum_{n=0}^{\infty}\eta_n(s_i,a_i)=\infty \text{ and } \sum_{n=0}^{\infty}\eta_n^2(s_i,a_i)<\infty \text{ for all}(s_i,a_i)$$

则对所有的状态-行为对 (s_i,a_i)，当 $n\rightarrow\infty$ 时，由 Agent 强化学习算法产生的 Q 值序列 $\{Q(s_i,a_i)\}$ 以概率 1 收敛到最优值 $Q^*(s,a)$。

设随时间变化的学习因子如下[8]：

$$\eta_n(s_i,a_i)=\frac{\alpha}{\lambda+visits_n(s,a)} \tag{17}$$

其中 $visits_n(s,a)$ 表示 agent 观察的总次数，a 和 λ 是正常数。

显然，学习因子 $\eta_n(s_i,a_i)\in[0,1]$ 将随着观察总数 $visits_n$ 的增加而减少，从而逐渐降低 Q 值更新的程度，以使 Q 值序列 $\{Q(s_i,a_i)\}$ 收敛到正确的 Q 函数。

证明：根据假设每个状态-行为对无限频繁地发生，对连续区间，其中每个状态-行为对至少被执行一次。设 $\hat{Q}_n$ 是 Q_n 的估计，则我们仅需证明最大误差在每个连续的区间内按折扣因子 β 减少 $(0<\beta<1)$。

令 ΔE 是这个估计的最大误差，即

$$\Delta E\equiv\max_{s,a}|\hat{Q}_n(s,a)-Q(s,a)| \tag{18}$$

则在第 $n+1$ 次迭代中更新的任意 Q_n，其估计 $\hat{Q}_{n+1}(s,a)$ 的误差可以定义如下：

$$\begin{aligned}|\hat{Q}_{n+1}(s,a)-Q(s,a)|&=|(r+\beta\max_{a'}\hat{Q}_n(s',a'))-(r+\beta\max_{a'}Q(s',a'))|\\&=\beta|\max_{a'}\hat{Q}_n(s',a')|-\max_{a'}Q(s',a')|\\&\leqslant\beta\max_{a'}|\hat{Q}_n(s',a')-Q(s',a')|\\&\leqslant\beta\max_{s',a'}|\hat{Q}_n(s'',a')-Q(s'',a')|\\&\leqslant\beta\Delta E\end{aligned}$$

$\because \Delta E\equiv\max_{s,a}|\hat{Q}_n(s,a)-Q(s,a)|, |\hat{Q}_n(s,a)-Q(s,a)|\leqslant\Delta E$

所以对所有的状态-行为对 (s_i,a_i)，更新后的 $\hat{Q}_{n+1}(s,a)$ 的误差最多为 $\hat{Q}_n$ 中最大误差 ΔE 的 β 倍。由于对任意状态-行为对 (s_i,a_i)，初始值 $\hat{Q}_0(s,a)$ and $Q(s,a)$ 是有界的，因此初始误差 ΔE_0 也是有界的。在所有 (s_i,a_i) 都被访问过的第一个区间内，其最大误差最多为 $\beta\Delta E_0$，故在 k 个区间后，误差最多为 $\beta^k\Delta E_0$。因为每一个状态-行为空间都是被无限频繁地访问，这样的区间也是无限的，所以当 $n\rightarrow\infty$ 时 $\Delta E_n\rightarrow 0$。定理得证。

注 1，在证明过程中我们用到了下面的引理：

引理 对任意函数 f_1 和 f_2，下面的不等式成立：

$$|\max_a f_1(a)-\max_a f_2(a)|\leqslant\max_a|f_1(a)-f_2(a)|$$

注 2，在上面的证明过程中我们引入了一个新的变量 s''，而且通过这个变量去最大化 Q 值。当附加的变量允许改变时，此最大值只可能更大或者至少相等。而且通过新变量的引入，我们获得了一个与 ΔE_n 的定义相匹配的表达式。

4 实验仿真与分析

将本文的研究结果应用于基于多 Agent 的电子商务交易系统(Multi-Agent Electronic Commerce Trade-off System, MAECTS)中。电子商务交易模型是在一定限制策略下的基于 Web 的交易方式，它在某种程度上满足了交易各方的利益需求，同时也通过一些规则规范了参与交易各方的交易行为，从而保证了交易效率和交易的有效性、公平性。

在实际系统中，MAECTS 中的交易环境通常是由多智能主体构成的虚拟市场，这种虚拟市场的状态空间往往是连续的和高维的。通过基于神经网络的多 Agent 强化学习，实现状态和行为空间的值函数映射，可以降低 Agent 在状态-行为空间计算的维度，优化 Agent 的行为策略，提高在复杂环境中各 Agent 逼近和预测未来交易行为以及未来市场状态趋势的能力和效率。

将基于神经网络的 Agent 强化学习模型作为 Agent 结构中的学习器，如图 3 所示。其中，S 为状态感知器，A 为行为选择器，ANNRL 为基于神经网络的强化学习器，E 为环境状态。

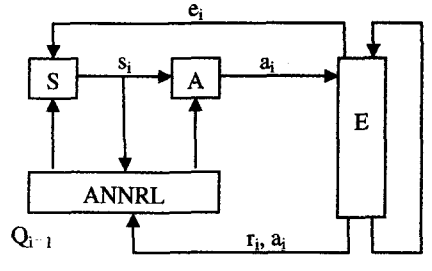

图 3 基于神经网络的 Agent 强化学习结构

(下转第 297 页)

7 Mesnier M, Ganger G R, Riedel E. Object-based storage. Communications Magazine, IEEE 2003,34(8):84～90
8 Qin L, Feng D. Active Storage Framework for Object-based Storage Device. In: Proceedings of the 20th International Conference on Advanced Information Networking and Applications (AINA'06) 2006, 2:97～101
9 Technology I. SCSI Object-Based Storage Device Commands (OSD), working draft Project T10/1355-D,2004
10 John W, Richard G, Carl S, et al. The HP AutoRAID hierarchical storage system. ACM Trans Computer System, 1996, 14(1): 108～136
11 Peter M C, David A P. Maximizing performance in a striped disk array. In: Proceedings of the 17th annual international symposium on Computer Architecture. Seattle, Washington, United States: ACM Press,1990
12 Reed D. Striping in aRAID Level 5 Disk Array Performance Evaluation Review. A quarterly publication of the Special Interest Committee on Measurement and Evaluation,1995,23(1):136
13 Peter M C, Lee E K. Striping in a RAID Level 5 Disk Array. ACM SIGMETRICS Performance Evaluation Review, 1995, 23(1):136～145
14 Yan X, Han J, Afshar R. CloSpan: Mining closed sequential patterns in large datasets. In: Proc. 2003 SIAM intConfData Mining (SDM'03),2003

(上接第 158 页)

实验中,对 Agent 的竞价行为进行仿真实验,并将实验结果与回归神经网络(RNN)及一般的强化学习算法(RL)相比较。

设 Agent$_i$ ($i=1,2,\cdots,N$)在参与某一商品交易的过程中对商品的报价是一个非平稳时间序列$\{Q_{i,j};i=1,2,\cdots,N;j=1,2,\cdots,M\}$,其中 N 为 Agent 个数,M 为竞价次数。

在实际应用中常常通过统计量来逼近模型以评价模型的性能。在本文中用均方根相对误差均值去评价算法 ANNRL。

当有 N 个 Agent 参与竞价时,为了比较基于神经网络的多 Agent 强化学习算法的性能,取 N 个 Agent l 步预测的 RMSRE 的平均值作为比较依据。其中均方根相对误差均值(the Mean of Root Mean Square Relative Error, MRMSRE)定义如下:

$$MRMSRE=\frac{1}{N}\sum_{j=1}^{N}\sqrt{\frac{1}{M}\sum_{j=1}^{m}\left(\frac{x_i-y_{j,l}}{x_i}\right)^2} \tag{19}$$

实验中取 10 个交易 Agent 对商品 U 在 157 次竞标中的报价,作为预测各交易主体在未来行动中的出价策略的数据基础。

实验结果如表 1 和图 4 所示,表中数据是分别对基于神经网络的多 Agent 强化学习算法(ANNRL)、RNN 和 RL 算法进行前 6 步预测所取得的均方根相对误差的平均值(MRMSRE)。

表 1 预测步数与均方根相对误差均值

STEP	ANNRL	RNN	RL
1	0.0015	0.0025	0.0033
2	0.0026	0.0387	0.0417
3	0.0562	0.0869	0.1169
4	0.1188	0.1321	0.2352
5	0.1297	0.2572	0.3061
6	0.1125	0.3829	0.4727

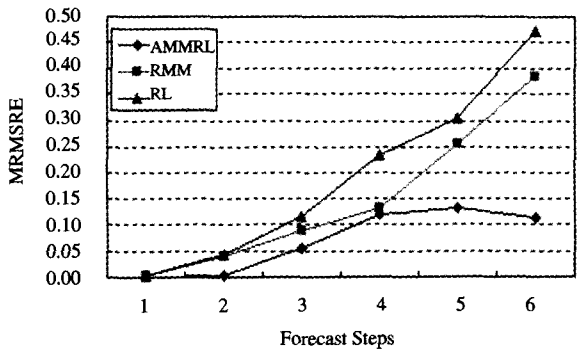

图 4 均方根相对误差均值比较效果图

实验结果表明,算法 ANNRL 在实际预测中具有非常理想的效果。在实验仿真中,仅仅 6 步预测即可获得 89.89% 的准确度,而且经过一定时间的学习后,ANNRL 算法收敛到一个极限值,这正是我们所期望的,因为每一个 Agent 是基于对其他主体和市场信息的信念而修正 Q 值的。

结论 本文设计了一个 Agent 强化学习的神经网络模型,实现了基于神经网络的 Agent 强化学习算法 ANNRL,证明了该算法在一定假设条件下收敛到最优 Q 值。由于神经网络可以逼近任意值函数,而通过值函数可以将高维空间映射到低维数据空间,因此 ANNRL 可以在一定程度上降低状态行为空间计算的 VC 维,同时通过随机梯度下降算法去最小化 Bellman 残差平方和,可以获得该算法更好的收敛速度和效果。我们在基于 Multi-Agent 的电子商务交易系统 MAECTS 中实现了 ANNRL 算法,通过对 MAECTS 中 Agent 的竞价行为进行仿真实验,验证了 ANNRL 算法具有良好的性能和行为逼近能力。

参 考 文 献

1 ZHANG Rubo, GU Guochang, et al. Reinforcement Learning Theory, Algorithm and Its Application [J]. Control Theory and Applications, 2000,(17)5: 637～641
2 Burnsa T R,Gomolinskab A. Socio-cognitive mechanisms of belief change Applications of generalized game theory to belief revision, social fabrication, and self-fulfilling prophesy [J]. Journal of Cognitive Systems Research,2001,2: 39～54
3 Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey [J]. Journal of Artificial Intelligence Research, 1996,4:237～285
4 Tsitsiklis J N, Roy B V. An analysis of temporal difference learning with function approximation [J]. IEEE Transactions on Automatic Control ,1997,42 (5) : 674～690
5 Watkins C J , Dayan P. Q-learning [J]. Machine Learning,1992, 8: 279～292
6 Singh S P,et al. Convergence results for single-step on policy reinforcement learning algorithms [J]. Machine Learning,2000,38 (3):287～308
7 Planning B C. Learning and coordination in multi-agent decision processes [J]. In: Proceedings of the Sixth Conference on Theoretical Aspects of Rationality and Knowledge [C]. Morgan Kaufmann,1996,195～210
8 Haykin S. NEURAL NETWORKS A Comprehensive Foundation [M]. In:Prentice Hall Tsinghua University Press, 2001
9 Sutton R S. Barto A G. Reinforcement Learning: An Introduction [M]. Cambridge, MA;MIT Press, 1998
10 ZHONG Yu,GU Guo-chang,ZHANG Ru-bo. Surey of distributed reinforcement learning algorithms in multi-agent systems [J]. Control Theory & Applications,2003,20(3):317～322
11 GAO Yang,CHEN Shi-fu,LU Xin. Research on Reinforcement Learning Technology: A Review. ACTA Automatica Sinica, 2004,30(1): 86～100
12 Taylor M E, Whiteson S, Stone P. Comparing Evolutionary and Temporal Difference Methods in a Reinforcement Learning Domain. In:Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2006), Seattle, WA July 2006. 1321～1328
13 Sherstov A A,Stone P. Improving Action Selection in MDP's via Knowledge Transfer. In:Proc AAAI-2005, Pittsburgh, USA