Dynamic Pricing by Multiagent Reinforcement Learning

Wei Han,Lingbo Liu,Huaili Zheng
Information Engineering College, Nanjing University of Finance and Economics, Nanjing, 210046, China

Abstract

Dynamic pricing in electronic marketplaces is a basic problem in electronic commercial. In multiagent environments, the optimal pricing policy of agent depends on the pricing policies of other agents. This makes the learning problem more problematic. This paper proposes an efficient online learning algorithm, which integrates the observed objective actions as well as the subjective inferential intention of the opponents. by establishing the decision model of other agents and predicting their proposed price in advance, agent becomes adaptive to its opponents and can make good decisions in long terms. The algorithm is proven to be effective when coming to the problem of seller's pricing in electronic marketplaces.

1. Introduction

Dynamic pricing in electronic marketplaces is a basic problem in electronic commercial. Diamond proposed a sequence decision model which is called BAZAAR [1], each pricing agent updates their belief about the reserve values of other agents by bayes method. Wu and Sun proposed a pricing strategy by using Gene Algorithm [2], the simulation results indicated that the best strategies usually is cooperative strategy. Kutschinskia *etc* let agent calculate it price by using reinforcement Q-learning. All of the above pricing algorithms did not predict the next actions of agent's opponents. That is to say, agents are supposed to be reactive rather than deliberative. Actually, the optimal police of any agent depends on polices of other agents in multiagent environments, which creates a situation of learning a moving and unclearly defined target. In order to obtain the knowledge of the environment and opponents, some recent works try to apply Reinforcement Learning (RL) to dynamic multiagent environments by integrating Markov Decision Process (MDP) and Game Theory (GM). The learning algorithm given by those works emphasizes both convergence and individual rationality. Bowling's algorithm[4] is the representative of this kind of work. It uses a variable learning rate, which increases when the expected payoff is smaller than the selected Nash equilibrium and vise versa. It is easy to see that the ultimate destination of this algorithm is the equilibrium of the game, which limits its application greatly. For example, in electronic marketplaces, sellers who sale the same product form a pricing game. According to analysis of economics, the best pricing policy is the equilibrium price under complete competitive market environment. However, actually no sellers will price like this. In fact, the pricing process is a dynamic adjustment between competitive sellers. So pricing between sellers is an online learning process rather than an offline learning result. Based on the pri cing problem in electronic marketplaces, this article contributes a new online multiagent learning method.

2. The pricing model

Electronic marketplaces are essentially multiagent systems, the self-interested seller agents interacts with each other through the market environments, which varies according to the supply and demand. The decision factors in sellers' pricing include cost, capacity, market demand function and the pricing polices of other sellers.

Definition 1. *the market demand function* is defined as an linear function of market average price level, that is $D(\bar{p}) = \max\{0, (q - h\bar{p})\}$, $q, h > 0$.

Definition 2. A seller agent is 3-tuple $Seller_i = (p_i, c_i, k_i)$, where p_i is the price of certain product, c_i is the cost, k_i is the production capacity.

Definition 3. A *electronic marketplaces* is described as $(n, s, A_{1.....n}, T, U_{1.....n})$, where n stands for the number of sellers who sell certain kind of homogenous product. $s = (p, c, k, q, h)$, $p = (p_1 ... p_n)$, $k = (c_1 ... c_n)$, $c = (c_1 ... c_n)$, q、h are parameters of market demand function. $A_i = \{a_1 .....a_{m_2}\}$ is the possible price set that agent i can offer. $A_{-i} = A_1 \times A_2 \times ... \times A_{i-1} \times A_{i+1} \times A_n$ is the set of joint actions of the other sellers. $T : S \times A \times S \to [0,1]$ is the transfer function of the market. $U_i : S \times A_i \to [0,+\infty]$ is the utility of agent i at present price.

Definition 4. The *pricing policy* of seller agent i is function $\pi : S \times A_i \to [0,1]$, which let agent choose a price stochastically according to market and his opponents.

Definition 5. The *utility function* of seller i is defined as $U_i(d, p_i, c_i, k_i) = (p_i - c_i) Y_i^{p_i}(d)$, where d is the demand of present market. $Y_i^{p_i}(d)$ is the quantity of products that agent i sold at price p_i.

978-0-7695-3258-5/08 $25.00 © 2008 IEEE
DOI 10.1109/ISECS.2008.179

We assume here that the products with lower price be sold first, that is buyers prefer cheap products.

3.Fictitious player learning

Q-learning algorithm is a reinforcement learning method for a single agent to learn optimal policy by exploring the whole state-action space in complex dynamic environments. When other agents adopt fixed stationary policies, multiagent environment turns to be a single-agent environment. The agent who adopts Q-learning algorithm will converge to the optimal policy [6].

The Opponent Modeling (OM) algorithm[7] is a variable type of standard Q-learning. The idea is to get explicit statistical models of the other players' actions, assuming that each opponont follow a stationary policy. In the algorithm, $c(s,a_{-i})/n(s)$ is the estimated distribution of the other agents choosing joint action a_{-i} based on their past observing history. The agent then plays the best response to this estimated distribution. The algorithm is essentially fictitious player adjustment process in game theory context, which has been proven to find equilibrium in certain type of games.

Algorithm 1. Fictitious player learning algorithm under undeterministic environments

1.Initialize $Q(s,a)$ arbitrarily.

$\forall s\in S, a_{-i}\in A_{-i}, c(s,a_{-i})\leftarrow 0$ and $n(s)\leftarrow 0$.

2. Repeat

a) From state s, select action a_i that maximizes,

$$\sum_{a_{-i}}\frac{c(s,a_{-i})}{n(s)}Q(s,<a_i,a_{-i}>).$$

b) Observing other agents actions a_{-i}, reward r, and next state s'.

$$c(s,a_{-i})\leftarrow c(s,a_{-i})+1;\ \ n(s)\leftarrow n(s)+1$$

$$Q(s,a)\leftarrow(1-\alpha)Q(s,a)+\alpha(r+\gamma V(s'))$$

Where $V(s)=\max_{a_i}\sum_{a_{-i}}\frac{c(s,a_{-i})}{n(s)}Q(s,<a_i,a_{-i}>)$.

4.Multiagent learning algorithm

Definition 6. Function $B_i^{o,t}(a_{-i}):A_{-i}\rightarrow R^+$ is called *objective belief revision function*, where $A_{-i}=A_1\times A_2\times...\times A_{i-1}\times A_{i+1}\times A_n$.

$$B_i^{o,t}(a_{-i})=B_i^{o,t-1}(a_{-i})+\begin{cases}1 & a_{-i}^{\,t-1}=a_{-i}^{\,t}\\ 0 & otherwise\end{cases}.$$

$P_i^{o,t}(a_{-i})=\dfrac{B_i^{o,t}(a_{-i})}{\sum_{a^{-i}}B_i^{o,t}(a_{-i})}$ is the combined distribution of the opponents' joint actions at period t.

An improvement of objective belief revision is to emphasize on recent observed actions of opponents. We call this exponent-index improvement.

Definition 7. Let agent i choose his action under FPL by taking the objective revised belief as his belief about the opponent agent j, we call this action the first-level-belief action about agent j.

Actually, the first-level-belief action about agent j is get by exchanging position with agent j, so it means the most possible action of agent j in the next period from viewpoint of agent i.

Definition 8. Let agent i choose his action under FPL by taking the $n-1$-level- belief action about agent j as the real action of agent j, we call this action the n-level-belief action of agent i about agent j.

Definition 9. If the first, second, …, n-level-belief action of agent i about agent j are $a_{j1}',a_{j2}',...,a_{jn}'$ respectively, then we call $P_{ij}^{s}:A_j\rightarrow[0,1]$ *subjective belief revision function*. n is the length of inference. $P_{ij}^{s}(a_{jk})=P_{ij}^{o,t}(a_{jk})+\Sigma_{jk}$, where $P_{ij}^{o,t}(a_{jk})$ is the marginal distribution of agent j ,which can be induced from the combined distribution of opponents' actions $P_i^{o,t}(a_{-i}):A_{-i}\rightarrow[0,1]$.

$\Sigma_{jk}=\sum_{p=1}^{n}\delta^{p}I(a_{jp}')$,

where $I(a_{jp}')=\begin{cases}1 & a_{jp}'=a_{jk}\\ 0 & otherwise\end{cases}$ is a label function, $\delta\in[0,0.5]$ is the believe level .

Algorithm 2. Multiagent Learning by predicting opponents' actions

1. Initialize Q value

$\forall s\in S, a_{-i}\in A_{-i}; c(s,a_{-i})\leftarrow 0; n(s)\leftarrow 0.$

2. Repeat

a). From present state s , observing the opponents' joint actions a_{-i},

$$c(s,a_{-i})\leftarrow\beta c(s,a_{-i})+1\ ;\ n(s)\leftarrow\beta n(s)+1\ ;$$

$$p(a_{-i})=c(s,a_{-i})/n(s)$$

b).For $\forall j \neq i$, computing marginal distribution $pp(j)$ for agent j, which can be induced from $p(a_{-i})$.

c). computing belief actions at each level

d).Proceeding *subjective belief revision for each opponent* according to definition 9.

e).Computing the combined distribution $p\_subject(a_{-i})$.

f).Choose action a_i that maximize $\sum_{a_{-i}} p\_subject(a_{-i})Q(s,(a_i,a_{-i}))$,

g).Updating

$$Q(n,s,a) \leftarrow (1-\alpha)Q(n-1,s,a) + \alpha(U_i + \gamma \max_{a'} Q(n,s',a')$$

h). Observing the new state s', $s \leftarrow s'$

5. Simulations

We assume there are 3 agents in electronic marketplace, vectors p,k,c,u. respectively stand for the price, capacity, cost and utility of the 3 agents. We also assume the market demand is linear function of average price, that is to say $D(\overline{p}) = \max\{0,(q-h\overline{p})\}$, q,h >0. Table 1 is the theoretical analysis results and table 2 is the results of learning algorithm. By comparing the results, we can draw the conclusion that multiagent learning algorithm enables seller agents possess some coordination ability, which makes them acquire greater utility. Seller agents also show some intelligence on the question of whether competes or coordinate. Sellers who have obvious dominance in competition will ally to eliminate those disadvantaged sellers (table 2, row 4 column 2). If their competition ability is near to each other, they will compete to some extent and then comprise with each other to avoid malignant competition. The willingness for competition of seller agents is relevant to market situations of supply and demand. Under situations that supply exceeds demand, competition between sellers is fierce and the result is near to equilibrium (table 2, row 2 column 2), and more sellers are eliminated (table 2, column 1). Under situations of demand exceeds supply, sellers will take advantage of cooperation to acquire greater benefit (table 2, column 2). Fig. 2 gives the utility curve of the first agent under situation of C=(10,10,18), K=(40,40,30), D=max{(100-P),0} at the first 100 period. The small vibration is because agent can only offer discrete possible price.

Table 1 The theoretical results under different market response functions

	D=max{(77-p),0}	D=max{(115-p),0)	D=max{(100-P),0}
C=(11,21,21)	P=(21,22,22)	no equilibrium	P=(23,24,24)
K=(44,23,23)	U=(440,5.5,5.5)		U=(528,48,48)
C=(16,16,25)	no equilibrium	P=(31,31,31)	no equilibrium
K=(34,34,22)			
C=(7,12,17)	no equilibrium	no equilibrium	no equilibrium
K=(45,26,19)			
C=(10,10,18)	no equilibrium	no equilibrium	P=(18,18,19)
K=(40,40,30)			U=(320,320,1)
C=(10,12,14)	no equilibrium	no equilibrium	no equilibrium
K=(40,30,20)			

Table 2 Average results of learning algorithm under different market respond functions

	D=max{(77-p),0}	D={(115-p),0)	D={(100-P),0}
C=(11,21,21)	P=(49,*[1],49)	P=(50,50,49)	P=(38,37,36)
K=(44,23,23)	U=(1064,0,410)	U=(1756,645,644)	U=(1088,368,345)
C=(16,16,25)	P=(29,28,*)	P=(47,46,46)	P=(28,28,28)
K=(34,34,22)	U=(442,408,0)	U=(1048,1020,462)	U=(408,408,53)
C=(7,12,17)	P=(31,30,*)	P=(32,32,53)	P=(24,23,23)
K=(45,26,19)	U=(1080,468,0)	U=(1125,520,272)	U=(765,286,114)
C=(10,10,18)	P=(32,31,*)	P=(38,37,37)	P=(21,21,21)
K=(40,40,30)	U=(880,840,0)	U=(1120,1180,553)	U=(438,433,65)
C=(13,13,13)	P=(16,16,16)	P=(49,48,48)	P=(30,29,29)
K=(30,30,30)	U=(61,61,61)	U=(1080,1050,1050)	U=(510,480,480)

* stands for the agent is eliminated

Acknowledgement

It is a project supported by Natural Science of Jiangsu Province (07KJD520070) and Nanjing University of Finances and Economics (C0728).

References

[1]. P.Diamond(1971), "A model of price adjustment", *Journal of Economics Theory*,vol.3,No.1 pp.156-168.

[2]. D.J.Wu, Yanjun Sun(2002). "Cooperation in multi-agent bidding". *Decision Support Systems*.Vol.33.pp.335-347

[3]. Erich Kutschinskia, Thomas Uthmannb, Daniel Polanic(2003). "Learning competitive pricing strategies by multiagents reinforcement learning". *Journal of Economic Dynamics & Control* .NO.27 ,pp.2207-2218.

[4]. Bowling, M(2002). "multiagent Learning Using a Variable Learning Rate".*Artificial Intelligence* Vol. 136. 215-250.

[5]. Fudenberg, D. *Learning of Game Theory*.University of Chinese People Press, .Beijing .2002(in Chinese).

[6]. Claus,C. Boutilier, C(2001). "The Dynamics of Reinforcement Learning in Cooperative Multiagent systems". *In: Wiliams: Proceeding of 18th International Conference on Machine Learning*. Word Science Press, MA,pp.27-34.

[7]. Wyatt,J(1997)."Exploration and Inference in Learning From Reinforcement. *Ph. D. Thesis*, Department of Artificial Intelligence, University of Edinburgh,.UK.pp.37-39.

[8]. Fudenberg,D. Levine,D.K. *The Theory of Learning in Games*. MIT Press,Cambridge MA,1960.

[9]. Kaelbling, L.P. Littman, M.L. Moore, A.W(1996). "Reinforcement learning: A survey". *Journal of Artificial Intelligence*.Vol. 4,pp. 237-285.

[10]. Filar, J. Vrieze, K. *Competitive Markov Decision Processes*. Springer press,New York,1997.