文章编号: 1001-0920(2005)12-1404-04

基于Q 学习的供应链分销系统最优订货策略研究

李随成[1,2], 尹洪英[2]
(1. 西北工业大学 管理学院, 西安 710072; 2. 西安理工大学 管理学院, 西安 710048)

摘 要: 研究由一个制造商和多个分销商组成的分销系统的最优订货策略问题. 在外部顾客需求不断变化的情况下, 以不断提高分销系统双方合作绩效为目标, 基于Q 学习算法来确定每个分销商的最优订货批量. 实例结果表明, 在外部需求不断变化的条件下, 该算法能简便地解决供应链企业分销系统合作中的最优订货批量问题.

关键词: 供应链管理; 分销系统; Q 学习算法; 最优订货批量
中图分类号: F273 **文献标识码**: A

Optimum Order Strategies for Distribution System of Supply Chain Based on Q-learning

LI Sui-cheng[1,2], *YIN Hong-ying*[2]
(1. Management School, Northwestern Polytechnical University, Xi'an 710072, China; 2. School of Business Administration, Xi'an University of Technology, Xi'an 710048, China. Correspondent: LI Sui-cheng, E-mail: lisc@xaut.edu.cn)

Abstract: The optimum order strategies for the distribution system consisting of a single manufacturer and multi-distributors in the supply chain management is discussed. In the case of the varying demand of customers, and with an aim at improving the cooperative performance of the distribution system, the optimum order batch of each distributor is determined based on the Q-learning algorithm. An example shows that this algorithm can simply solve the problem of the optimum order batch in the distribution system in the case of external ongoing changes in demand.

Key words: Supply chain management; Distribution system; Q-learning algorithm; Optimum order batch

1 引 言

顾客的需求越来越决定供应链的存亡, 供应链的重心正逐渐转向需求方, 对供应链的研究也从最开始的供应与采购关系转向了分销系统. 分销系统中分销商对顾客的需求较为敏感, 它也更接近最终用户, 能够更好地掌握市场需求的变化. 制造商与分销商合作关系的主要特征就是二者及时保持市场信息的交流, 共同确定制造产量的多少与产品销售单价的高低, 使制造商的产量与分销商的订单数量保持一致, 否则有可能出现两次订单的情况[1]. 分销网络优化已引起人们的广泛关注, 许多文献针对不同的背景、目标, 通过选择不同的方案设定不同的供应链管理的外部环境来分析分销网络. 文献[2]研究由一个供应商、一个分销中心和若干零售商组成的两级分销系统, 其目标是使整个供应链中的总成本最小, 其中假设零售商处的需求所服从的分布是其分布函数的卷积为封闭的一类函数, 由此计算最优的订货策略. 文献[3]针对由多个制造商和多个分销商组成的分销网络, 在综合考虑各类成本(库存成本、订货成本、运输成本和缺货成本)的基础上, 以分销网络的物流成本最小化为目标, 将服务水平作为约束条件进行建模, 由此得出最优订货批量.

以上文献从不同的角度探讨了分销系统的最优

收稿日期: 2004-11-09; **修回日期**: 2005-04-07.
基金项目: 陕西省自然科学基金项目(04JK263).
作者简介: 李随成(1962—), 男, 河南孟州人, 教授, 从事物流与供应链管理等研究; 尹洪英(1979—), 女, 山东鄄城人, 硕士生, 从事物流与供应链管理的研究.

订货策略, 通过建立数学模型来分析问题, 而且均基于年需求量已知的假设条件. 鉴于当前顾客需求的不确定性越来越大, 年与年之间的顾客需求存在着很大的差异, 本文研究由一个制造商和多个分销商组成的分销系统在顾客需求不断变化的条件下, 以不断提高合作双方的合作绩效为目标, 寻求最优订货策略的方法. 提出了在供应链企业分销系统中不断提高双方合作绩效的基础上, 用Q 学习算法来确定最优订货批量. 实例证明, 这种算法能够很简便地在外部需求不断变化的条件下, 解决供应链企业分销系统合作中的最优订货批量的确定问题.

2　强化学习下的Q 学习算法

强化学习(RL)是一种机器学习方法, 它是一个试错的过程[4], 是求解随机的、序贯的Markov 决策过程(MDPs)的有效方法, 在机器维护、库存控制等决策优化问题中有着广泛的应用[5], 其基本原理见文献[6]. 它的主要优点是能够在学习中进步以作出好的决策[7], 通常用MDPs 作为基础数学模型. 一个有限的MDPs 由 4 元组$\langle S, A, p, r\rangle$来定义, 其中: S 是有限的状态集, A 是有限的行动集, p 是转移概率, r 是即时报酬函数. 基本学习机制如图 1 所示[8].

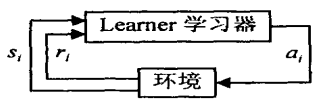

图 1　强化学习基本学习机制

累计折扣回报的期望值定义为

$$Q^{\pi}(s,a) = E\left\{\sum_{t=0}^{\infty}\gamma r_t \,\middle|\, s_t = s, a_t = a, \pi\right\}. \quad (1)$$

上式表示在状态s 和行动a 选择策略π时的期望累计折扣报酬. 其中$0 \le \gamma \le 1$为折扣因子, 表示较近的回报比较远的回报更重要.

最优价值函数定义为

$$Q^{*}(s,a) = \max_{\pi} Q^{\pi}(s,a). \quad (2)$$

由Q^{*}所确定的策略$\pi = \arg\max Q^{*}(s,a)$即为最优策略$\pi^{*}$. 强化学习算法有基于模型和无模型两种途径. 学习最优值函数Q 学习算法是Watkins[9] 提出的一种无模型强化学习算法. 它用状态s 下采取行动α的下一个状态s'对假定行动α'所对应的最大Q值更新当前的$Q(s,\alpha)$, 其更新公式见文献[10].

3　供应链分销系统最优订货批量问题描述

在此构建由一个制造商和多个分销商组成的供应链分销系统. 在该分销系统中, 制造商与分销商之间建立良好的合作关系, 双方最终目的都是为了最大程度地满足顾客需求, 并使双方合作绩效不断提高. 其中分销商把所确定的需求数量告诉制造商, 双方经过协商确定最优订货批量, 并对合作绩效进行评价. 其示意图如图 2 所示.

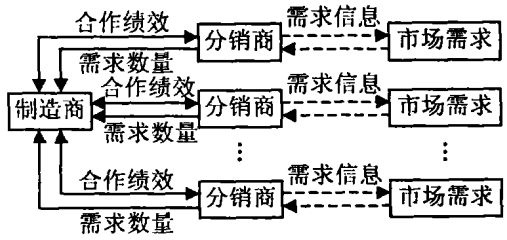

图 2　分销系统合作示意

在应用Q 学习算法求解问题时, 首先需将状态和决策空间离散化, 然后建立一个信息不完全的有限马尔可夫决策过程. 因为分销活动本身就是许多离散的阶段决策过程, 其中每一次制造商对分销商$i(i = 1, 2, \cdots, n)$的补货就是一个离散阶段, 所以它本身就是一个马尔可夫决策过程.

设每个分销商i 对制造商的最优订货量为$C(i)$, 总订货量$\mathrm{Total} = \sum_{i=1}^{n} C(i)$, 变量$S_y(i)$代表分销商$i$的订货量为$y$ 时占总订货量的百分比, 即$C(i) = y$, 所以

$$S_y(i) = \frac{C(i)}{\sum_{i=1}^{n} C(i)}. \quad (3)$$

在合作开始时间t, 分销商i 与制造商协商确定其初始订货批量为

$$C(i,t) = y,$$
$$S_y(i,t) = \max[S_j(i,t) + \eta(i,j,t)], \quad (4)$$

其中: j 代表以前合作中的订货批量, η是一个整数均匀分布的随机变量.

4　供应链企业中分销系统最优订货策略的Q 学习解法

对于已给定的环境, 学习任务是在外部需求不断变化的情况下, 找到在合作中最优的订货批量, 如果它使得整体合作绩效达到最优, 那么这个值就是最优的. 在合作初始阶段, 为了较快地接近最优值, 可以利用最近的历史销售数据来估计需求数量, 以此来确定第 1 次的订货批量. 设此分销系统在第t次合作结束时制造商与分销商i 的合作绩效的评价值为$N(i,t)$, 期望值为$\hat{N}(i,t)$. 期望值$\hat{N}(i,t)$ 对应于过去合作绩效的轨迹并作为当前绩效的一个基准, 以此来估计它是否进步了. 如果$N(i,t) > \hat{N}(i,t)$, 那么合作者将被奖励; 否则他们将受到惩罚. 如果在第t次合作中分销商i 的订货批量为y (即$C(i,t) = y$), 那么双方将按以下步骤修正其订货批量:

Step 1: 初始化$\hat{N}(i,1)$.

在第 1 次分销系统合作结束时, 即$t = 1$时, 对

制造商与分销商 $i=1$ 时的合作绩效进行评价(合作绩效可从双方对外部环境的适应程度来评价),给出供应链分销系统中双方第 1 次合作的合作绩效值 $N(i,1)$,并将第 1 次合作期望值 $\hat{N}(i,1)$ 初始化为 $\hat{N}(i,1)=N(i,1)$.

Step 2: 计算 $\hat{N}(i,t)$ 的值

令 $t=t+1$,在 t 结束时,对合作结果进行评价,给出供应链分销系统中双方第 t 次合作的合作绩效值 $N(i,t)$,并计算 $\hat{N}(i,t)$ 的值:

$$\hat{N}(i,t)=\lambda N(i,t-1)+(1-\lambda)\hat{N}(i,t-1), \tag{5}$$

其中 $0<\lambda<1$ 为一恒量,它度量过去的合作绩效对当前合作绩效的影响

Step 3: 计算合作绩效变化率 $r(i,t)$ 的值:

$$r(i,t)=\frac{N(i,t)-\hat{N}(i,t)}{\hat{N}(i,t)}. \tag{6}$$

Step 4: 计算供应链分销系统合作中最优订货批量百分比的修正值 $S_y(i,t+1)$:

$$S_y(i,t+1)=S_y(i,t)+\alpha r(i,t), \tag{7}$$

其中 α 是一个决定学习速度的正恒量,一般取 $\alpha=0.1$,它已被实验证明是可取的,尤其在动态环境下[7].

Step 5: 计算供应商 i 的最优订货批量的修正值 $C(i,t+1)$:

$$C(i,t+1)=\mathrm{Total}\times S_y(i,t+1). \tag{8}$$

在下一次合作时,分销商 i 以 $C(i,t+1)$ 作为最优订货批量进行订货 返回 Step 2,同时令 $i=i+1$(直到 $i=n$ 结束). 返回 Step 1.

经过上述步骤的循环能使供应链分销系统在外部顾客需求不断变化的情况下,使制造商与分销商之间的合作不断优化 需要说明的是:因为只有在第 1 次合作结束时 $\hat{N}(i,1)$ 才能初始化为 $\hat{N}(i,1)=N(i,1)$,所以分销系统在第 2 次合作结束时才开始学习.

5 应用实例

下面给出一个比较简单的实际应用例子. 由一个制造商和 3 个分销商组成供应链企业的分销系统,它们之间建立了良好的合作伙伴关系 该制造商生产一种较畅销的饮料,饮料的需求随着季节和外包装的变化而影响顾客需求数量的变化,现实中以一周为订货周期来订货,现需要确定其各个分销商的最优订货批量 假设根据上周的历史销售数据来初步确定初始订货批量,第 1 次合作与第 2 次合作的绩效评价值如表 1 所示,现在对其进行优化 这里取 $\alpha=0.1$, $\lambda=0.3$.

表 1 各分销商初始订货批量及第 1 次合作与第 2 次合作的绩效评价值

分销商代码	初始订货批量/个	第 1 次合作绩效评价值	第 2 次合作绩效评价值
A	218	0.75	0.83
B	245	0.79	0.76
C	337	0.82	0.88

以分销商 A 为例说明以上算法,其中 Total = 800, $S_y(A,2)=0.2725$. 首先在 Step 1 中得到 $N(A,1)=0.75$,并将第 1 次合作期望值 $\hat{N}(A,1)$ 初始化为 $\hat{N}(A,1)=N(A,1)=0.75$. 其次,在 Step 2 中 $t=2$,在第 2 次合作结束时双方的合作绩效值 $N(A,2)=0.83$,根据式(5) 计算 $\hat{N}(A,2)$ 值:

$$\hat{N}(A,2)=\lambda N(A,1)+(1-\lambda)\hat{N}(A,1)=0.3\times0.75+(1-0.3)\times0.75=0.75$$

(在第 2 次合作时 $\hat{N}(A,2)=N(A,1)$,在下一次合作中,即 $t=3$ 时,$\hat{N}(A,3)$ 一般不等于 $N(A,2)$). 然后根据式(6) 计算合作绩效变化率

$$r(A,2)=\frac{N(A,2)-\hat{N}(A,2)}{\hat{N}(A,2)}=\frac{0.83-0.75}{0.75}\approx0.107.$$

再根据式(7) 计算供应链分销系统合作中最优订货批量的修正值

$$S_y(A,3)=S_y(A,2)+\alpha r(A,2)=0.2725+0.1\times0.107=0.2832$$

最后根据式(8) 计算供应商 A 的最优订货批量的修正值

$$C(A,3)=\mathrm{Total}\times S_y(A,3)=800\times0.2832=226.56\approx227\text{(个)}.$$

在下一次合作时,A 将以 227 作为最优订货批量进行订货 返回 Step 2.

同理,按照上述步骤可求得修正后分销商 B 和 C 的订货批量值分别为 242 个和 343 个,总订货量扩张到 227 + 242 + 343 = 812(个),这是分销系统内部强化学习和市场外部利好信息相互影响的结果 接下来,双方以这些值作为其下次的订货批量进行合作 下一次合作结束后再次得到其合作的绩效评价值,再对订货批量进行修正,一直这样反复下去

6 结 语

本文探讨了由一个制造商和多个分销商组成的分销系统,在外部顾客需求不确定的情况下,双方建立良好的合作关系后,为达到其整体合作绩效最优,用 Q 学习算法来确定各个分销商的最优订货批量 该方法比较简单,而且对需求的变化可以作出快速反应,但需要合作双方一直保持良好的合作关系,

且需要一定的学习时间

参考文献(References)

[1] 张文慧, 赵道致. 供应链中合作机制的研究[J]. *天津理工学院学报*, 2002, 18(4): 108-111.
(Zhang W H, Zhao D Z. Study of Cooperation Mechanisms in Supply Chain[J]. *J of Tianjin Institute of Technology*, 2002, 18(4): 108-111.)

[2] 张钦, 沈厚才, 达庆利. 供应链管理: 两级分销系统的最优订货策略[J]. *系统工程学报*, 2002, 17(4): 303-308.
(Zhang Q, Shen H C, Da Q L. Supply Chain Management: Optimizing Order Strategies of Two-echelon Distribution System [J]. *J of Systems Engineering*, 2002, 17(4): 303-308.)

[3] 王迎军, 高峻峻. 供应链分销系统优化及仿真[J]. *管理科学学报*, 2002, 5(5): 79-84.
(Wang Y J, Gao J J. Optimization and Simulation of Distribute Systems in a Supply Chain [J]. *J of Management Science in Chain*, 2002, 5(5): 79-84.)

[4] Antonio Murciano, Jose del R Millan, Javier Zamora. Specialization in Multi-agent Systems Through Learning[J]. *Biological Cybernetics*, 1997: 76(5): 375-382.

[5] 李春贵, 刘永信. 一种有限时段Markov 决策过程的强化学习算法[J]. *广西工学院学报*, 2003, 14(1): 1-4.
(Li C G, Liu Y X. An Algorithm of Reinforcement Learning for Finite-horizon Markov Decision Processes [J]. *J of Guangxi University of Technology*, 2003, 14(1): 1-4.)

[6] 王醒策, 张汝波, 顾国昌. 多机器人动态编队的强化学习算法研究[J]. *计算机研究与发展*, 2003, 40(10): 1444-1450.
(Wand X C, Zhang R B, Gu G C. Research on Dynamic Team Formation of Multi-robots Reinforcement Learning [J]. *J of Computer Research and Development*, 2003, 40(10): 1444-1450.)

[7] Kim C O, Jun J, Baek J K, et al. Adaptive Inventory Control Models for Supply Chain Management[J]. *Int J of Advanced Manufacturing Technology*, 2004, 26(7): 1184-1192.

[8] 张春阳, 陈小平, 刘贵全, 等. Q-learning 算法及其在囚徒困境问题中的实现[J]. *计算机工程与应用*, 2001, 13(1): 121-128.
(Zhang C Y, Chen X P, Liu G Q. Q-learning Algorithm and Its Usage in Prisoner's Dilemma[J]. *J of Computer Engineering and Application*, 2001, 13(1): 121-128.)

[9] 蒋国飞, 吴沧浦. Q 学习算法在库存控制中的应用[J]. *自动化学报*, 1999, 25(2): 236-241.
(Jiang G F, Wu C P. Inventory Control Using Q-learning [J]. *Acta Automatica Sinica*, 1999, 25(2): 236-241.)

[10] 李学勇, 欧阳柳波, 李国徽. 基于隐偏向信息学习的强化学习算法[J]. *南华大学学报(理工版)*, 2004, 18(2): 10-16.
(Li X Y, Ou Yang L B, Li G H. Reinforcement Learning Based on Hidden Biasing Information Learning [J]. *J of Nanhua University (Science & Engineering Edition)*, 2004, 18(2): 10-16.)

(上接第 1403 页)

[8] 张端金, 吴捷. 具有区域极点和方差约束的Delta 算子系统鲁棒 H_∞ 滤波[J]. *控制与决策*, 2004, 19(1): 12-16.
(Zhang D J, Wu J. Robust H_∞ Filtering for Delta Operator Formulated Systems with Circular Pole and Error Variance Constraints[J]. *Control and Decision*, 2004, 19(1): 12-16.)

[9] Guo Z. A Survey of Satisfying Control and Estimation [A]. *Proc of the 14th IFAC World Congress* [C]. Beijing, 1999: 443-447.

[10] Skelton R E, Iwasaki I. Liapunov and Covariance Controllers[J]. *Int J Control*, 1993, 57(3): 319-536.

[11] ChilaliM, Gahinet P. Design with Pole Placement Constraints: An LMI Approach[J]. *IEEE Trans on Automatic Control*, 1996, 41(3): 358-367.

[12] Petersen I R, Hollot C V. A Riccati Equation Approach to the Stabilization of Uncertain Linear Systems[J]. *Automatica*, 1986, 22(4): 397-411.

[13] Kyung-Soo Kim, Faryar Jabbari. Using Scales in the Multiobjective Approach [J]. *IEEE Trans on Automatic Control*, 2000, 45(5): 973-977.

[14] ChilaliM, Gahinet P, Apkarian P. Robust Pole Placement in LMI Regions [J]. *IEEE Trans on Automatic Control*, 1999, 44(12): 2257-2270.