第25卷 第2期
1999年3月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 25, No. 2
Mar., 1999

短 文

Q学习算法在库存控制中的应用[1)]

蒋国飞 吴沧浦

(北京理工大学自动控制系 北京 100081)

摘 要 Q学习算法是Watkins提出的求解信息不完全马尔可夫决策问题的一种强化学习方法.这里提出了一种新的探索策略,并将该策略和Q学习算法有效结合来求解一类典型的有连续状态和决策空间的库存控制问题.仿真表明,该方法所求解的控制策略和用值迭代法在模型已知的情况下所求得的最优策略非常逼近,从而证实了Q学习算法在一些系统模型未知的工程控制问题中的应用潜力.

关键词 Q学习,马尔可夫决策过程,库存控制,连续状态和决策空间,探索策略.

INVENTORY CONTROL USING *Q*-LEARNING

JIANG Guofei Wu Cangpu

(*Department of Automatic Control, Beijing Institute of Technology, Beijing* 100081)

Abstract *Q*-learning is a reinforcement learning method to solve Markovian decision problems with incomplete information. In this paper, we present a novel exploration strategy and use *Q*-learning method with this strategy to solve a typical inventory control problem with continuous state and decision space. Simulation results are included to show that the optimal policy given by *Q*-learning can well approximate to the accurate one.

Key words *Q*-learning, Markovian decision problem, inventory control, continuous state and decision space, exploration strategy.

1 引言

Q学习算法是求解信息不完全马尔可夫决策问题的一种有效的强化学习方法.自从Watkins[1]在1989年提出Q学习算法并证明了其收敛性后,该算法在强化学习研究领域中得到了人们的普遍关注.然而,目前对Q学习算法的应用研究还很不够,基本上限于人

1) 国家自然科学基金资助项目.

收稿日期 1997-04-14 收修改稿日期 1997-11-10

工智能方面,如最短路径搜索问题[1,2]、动物觅食游戏[3]等,实际应用的例子很少;另一方面对该算法的研究也只局限于一些离散状态和决策空间.迄今为止还没有见到在连续状态和决策空间上应用的例子,而实际应用中许多随机控制系统的状态和决策是连续的.本文通过离散化问题的状态和决策空间,提出用 Q 学习算法来求解运筹学中一类典型的有连续状态和决策空间的库存控制问题,同时针对决策空间离散化后可行控制数目较大的情况,提出了一种新的探索策略,来对控制和探索之间的冲突进行有效的折衷.

先简述有限马尔可夫决策问题的模型如下:在每个时间步 $k=1,2,\cdots$,控制器观察马氏过程的状态为 x_k,选择决策 a_k,收到即时报酬 r_k,并使系统转移到下一个状态 y_k,转移概率为 $P_{x_k,y_k}(a_k)$.控制目的是寻求一最优控制策略,使报酬函数 $E\left\{\sum_{i=0}^{\infty}\gamma^i r_{k+i}\right\}$ 最大,其中 $0\leqslant\gamma\leqslant 1$ 为折扣因子.

在一些实际例子中,状态转移概率和所获报酬是未知的,求解这类信息不完全的马氏决策问题有两种主要方法:贝叶斯方法和非贝叶斯方法.而非贝叶斯方法又分为直接法和非直接法.直接法是不估计系统模型而直接基于观察值来求解最优策略[2].Q 学习算法属于直接法.

给定一个策略 π,定义 Q 值为在状态 x,控制 a 及后续策略 π 下的报酬折扣和的期望.即

$$Q^{\pi}(x,a)=R(x,a)+\gamma\sum_{y}P_{xy}(a)V^{\pi}(y), \tag{1}$$

其中　$R(x,a)=E\{r\mid x,a\}$,$V^{\pi}(y)=\max\limits_{a}Q^{\pi}(y,a)$.$Q$ 学习就是要在转移概率和所获报酬未知的情况下估计最优策略的 Q 值.定义 $Q^{*}(x,a)\equiv Q^{\pi^{*}}(x,a)$,$\forall\ x,a$.其中 π^{*} 表示最优策略.

在线 Q 学习方法实现如下:在每个时间步 k,观察当前状态 x_k,选择和执行控制 a_k,再观察后继状态 y_k 及接受即时报酬 r_k,然后根据下式调整 Q_{k-1} 值:

$$Q_k(x,a)=\begin{cases}(1-\alpha_k)Q_{k-1}(x,a)+\alpha_k[r_k+\gamma V_{k-1}(y_k)], & \text{如果}(x,a)=(x_k,a_k),\\ Q_{k-1}(x,a), & \forall\ (x,a)\neq(x_k,a_k)\end{cases} \tag{2}$$

其中　α_k 为学习因子,$V_{k-1}(y)\equiv\max\limits_{b}\{Q_{k-1}(y,b)\}$.

Watkins[1]证明了学习因子序列 $\{\alpha_k\}$ 在满足一定的条件下,如果任一个 (x,a) 二元组能用方程(2)进行无穷多次迭代,则当 $k\to\infty$ 时,$Q_k(x,a)$ 以概率1收敛于 $Q^{*}(x,a)$.

2　库存控制及其有限马氏决策过程模型

设 x_k 是产品在 k 段开始时的库存量,u_k 是 k 段开始时的订货量(假设订了货即可立即交货),w_k 是 k 段中的需求量,是一个随机变量,R 是产品的仓库库存限量.$h(h\geqslant 0)$ 是单位产品的库存费用,$p(p>0)$ 是单位产品的脱需费用,$c(c>0)$ 是单位产品的订货费用.假定如需求量超出可能的供应量时,不允许先记帐而后补,即 x_k 不允许取负值.此时可建立库存优化控制问题的模型如下:

1)状态转移规律由下列方程确定

$$\begin{aligned} &x_{k+1} = \max(0, x_k + u_k - w_k), \\ &\text{约束:} x_k + u_k \leqslant R \quad k = 0,1,\cdots; \end{aligned} \tag{3}$$

2) $w_k(k=0,1,\cdots)$的概率分布相同且相互独立;

3) 控制的目的是选择订货策略 $u_k(x_k)$来极小化指标函数

$$\underset{\substack{w_k \\ k=0,1,\cdots}}{E} \left\{ \sum_{k=0}^{\infty} \gamma^k [C(u_k) + h \max(0, x_k + u_k - w_k) + p \max(0, w_k - x_k - u_k)] \right\}, \tag{4}$$

其中 γ 为折扣因子,$C(u)$为如下定义函数:

$$C(u) = \begin{cases} K + c \cdot u, & \text{如 } u > 0, \\ 0, & \text{如 } u = 0. \end{cases} \tag{5}$$

如果产品的仓库库存没有限量(即取消(3)式中的约束 $x_k+u_k\leqslant R$),Bertsekas 在文献[4]中证明了当 $p>c$ 时可求得以上问题的最优策略为

$$u^*(x) = \begin{cases} S^* - x, & \text{如 } x < s^*, \\ 0, & \text{其他.} \end{cases} \tag{6}$$

其中 S^* 为函数 $G^*(y)=cy+L(y)+\underset{w}{E}\{J^*(y-w)\}$的极小点,而 $L(y)=p\underset{w}{E}\{\max(0, w-y)\}+h\underset{w}{E}\{\max(0,y-w)\}$. s^* 为满足 $G(y)=K+G(S^*)$的 y 的最小值.

用 Q 学习算法求解马氏决策问题,不须知道过程的状态转移规律,只要在决策过程中能观测到系统的四元组序列$(x_k,u_k,r(x_k,u_k),y_k)$(其中 x_k 为当前状态,u_k 为当前控制,$r(x_k,u_k)$为在系统状态 x_k 上执行控制 u_k 后得到的报酬,y_k 为后继状态.),就能通过学习求得问题的最优策略. 在仿真中,以上所建立的库存控制模型代表实际过程向 Q 学习提供这个四元组序列. 而在实际应用中,在线 Q 学习算法可以直接从系统观测这个四元组序列.

要应用 Q 学习方法来求解以上有连续状态和决策空间的库存控制问题,首先需将状态和决策空间离散化后建立一个信息不完全的有限马尔可夫决策过程. 设将状态离散化成 N 段,$x=0,1,\cdots,i,\cdots,j,\cdots,N$;将控制离散化成 M 段,$u=0,1,\cdots,k,\cdots,M$;格子点间距为 L. w_k 为一个连续的随机变量,不需要离散化. 这样就可以推出即时报酬 $r(i,k)=C(k\cdot L)+h\ \max(0,i\cdot L+k\cdot L-w_k)+p\ \max(0,w_k-i\cdot L-k\cdot L)$;状态转移概率 $P_{ij}(k)=P(j\cdot L\leqslant\max(0,i\cdot L+k\cdot L-w_k)<(j+1)\cdot L)$. 然而由于随机需求量 w_k 的概率分布是未知的,所以式中的 $P_{ij}(k)$实际上也是未知的. 在仿真中系统状态根据式(3)进行转移并集结到相应的格子点上. 在决策空间离散化后,为了加快 Q 学习算法的收敛速度和求得最优策略,下面提出了一种新的探索策略来选择控制行为.

3 探索策略

在 Q 学习方法的实现中,有多种探索方法[5]. 但最为常用的是 Boltzmann 分布探索. 假定在时间步 k 时,状态为 x_k,控制 $a\in A(x_k)$,$A(x_k)$是状态为 x_k 时的控制集合. 定义 $E(a)=Q(x_k,a)$,$\forall\ a\in A(x_k)$. 则控制器用以下概率来选择控制 a

$$P(a) = \frac{e^{E(a)/T}}{\sum_{b\in A(x_k)} e^{E(b)/T}}, \tag{7}$$

其中 T 为温度因子,随着学习进行而逐渐衰减.

根据 Boltzmann 分布探索方法的特点,在其基础上提出了一种新的基于计数器的直接探索方法. 在每个时间步 k,控制器用以下概率来选择控制 a

$$P(a) = (1 - \Gamma) \cdot P_{\text{exp lore}}(a) + \Gamma \cdot P_{\text{exp loit}}(a), \tag{8}$$

其中 $P_{\text{exp loit}}(a)$为 Boltzmann 分布探索,如式(7)所示. 式(8)中的 $\Gamma(0<\Gamma<1)$选一固定值或一简单的分段函数. 而式(8)中的 $P_{\text{exp lore}}(a)$由下式确定

$$P_{\text{exp lore}}(a) = \frac{P_{\text{exp lore}}(a)}{\sum_{b \in A(x_k)} P_{\text{exp lore}}(b)}, \tag{9}$$

其中

$$E_{\text{exp lore}}(a) = \frac{\alpha\beta}{(\alpha + n(x_k,a)^m)}, \tag{10}$$

(10)式中 $n(x_k,a)$为到时间步 k 为止,系统在状态 x_k 执行控制 a_k 的次数. α,β 为常数且满足 $\alpha>0,\beta>0$. 参数 $m(m>0)$随学习进行而逐渐衰减到零. 由于$0\leqslant P_{\text{exp lore}}(a)\leqslant 1$,$0\leqslant P_{\text{exp loit}}(a)\leqslant 1$,$\sum_{a\in A(x_k)} P_{\text{exp lore}}(a)\equiv 1$,$\sum_{a\in A(x_k)} P_{\text{exp loit}}(a)\equiv 1$,由(8)式显然可证$0\leqslant P(a)\leqslant 1$及 $\sum_{a\in A(x_k)} P(a)\equiv 1$. 在学习过程的初始阶段,对$\forall\ a\in A(x_k)$,式(7)中的温度 T 和式(10)中的参数 m 越大,$P_{\text{exp loit}}(a)$分布越均匀而 $P_{\text{exp lore}}(a)$的分布越有差异,此时系统将主要根据 $P_{\text{exp lore}}(a)$来选择控制行为,从而保证环境被充分探索和获取各种控制经验. 随着学习的进行,T,m 逐渐衰减,$\forall\ a\in A(x_k)$,$E_{\text{exp lore}}(a)$的值将逐渐趋向一致,从而使 $E_{\text{exp loit}}(a)$分布越有差异而 $E_{\text{exp lore}}(a)$的分布越为均匀,此时代表控制经验的 $E_{\text{exp loit}}$在控制行为 a 的选择中逐渐增加了作用而代表环境探索的 $E_{\text{exp lore}}$逐渐减少作用,从而增加已有控制经验的利用和改善系统控制性能. 随着学习的进一步进行,m,T 将趋向于零,$\forall\ a\in A(x_k)$,$\lim_{m\to 0} E_{\text{exp lore}}(a)=\frac{\alpha\beta}{(\alpha+1)}$,$\lim_{m\to 0} P_{\text{exp lore}}(a)=\frac{1}{n_k}$,$n_k$ 为集合 $A(x_k)$中控制的个数,此时系统将主要根据 $P_{\text{exp loit}}(a)$来选择控制行为,从而来充分利用控制经验和保证算法的收敛速度. 在 m,T 都趋于零时探索策略实质上已变成一半均匀分布探索,即以概率值为 Γ 选择对应 Q 值最大的控制,而以概率值为$1-\Gamma$ 随机均匀地在所有可行控制中进行选择,这样在学习过程的最后阶段一方面能保证算法的收敛速度,另一方面也仍然保留了对环境的一定探索以求得最优策略.

4　仿真及结果

在仿真实例中,产品的仓库库存限量 $R=100.0$,单位产品的库存费用 $h=1.0$,单位产品的脱需费用 $p=9.0$,单位产品的订货费用 $c=2.0$,订货固定费用 $K=0.0$. 折扣因子 $\gamma=0.90$. 随机需求 w_k 服从$(50,10^2)$的正态分布. 在模型已知时,将状态、控制分别离散化成 $N=M=50$段,格子点间距为 $L=2.0$,用值迭代法求得问题的最优策略如图1所示. 同时在模型未知时,将状态、控制分别离散化成 $N=M=25$段,格子点间距为 $L=4.0$. Q 学习方法结合新的基于计数器的探索方法在经过15.9万次迭代后求得问题的控制策略如图2所示,继续迭代控制策略可收敛到图1所示的最优策略. 而 Q 学习方法结合 Boltzmann

分布探索在经过32.4万次迭代后求得问题的控制策略如图3所示，但该方法在经过上千万次迭代后仍然不能完全收敛到问题的最优策略. 图1－3中 $u(x)$ 表示控制策略，x 表示状态.

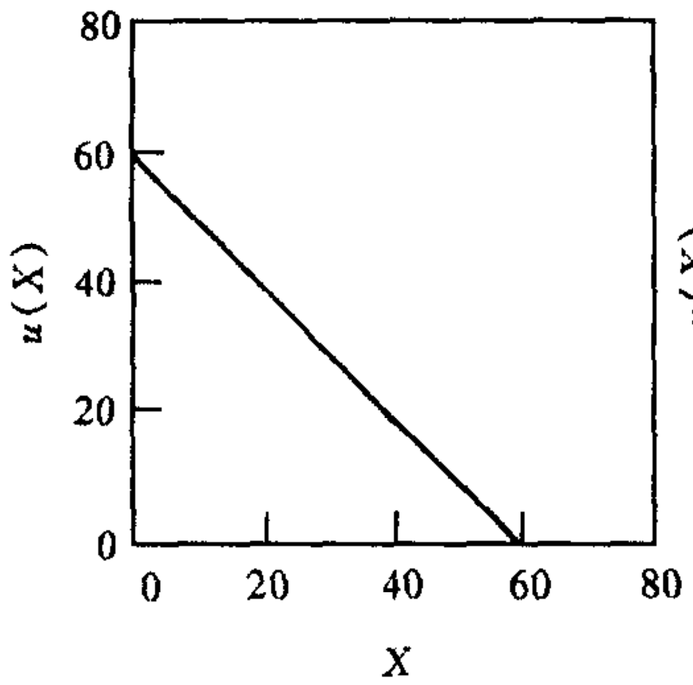

图1

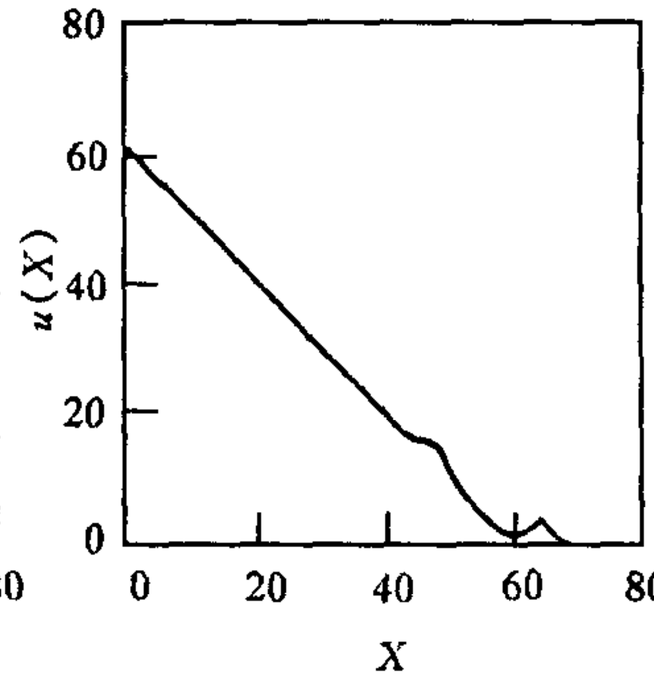

图2

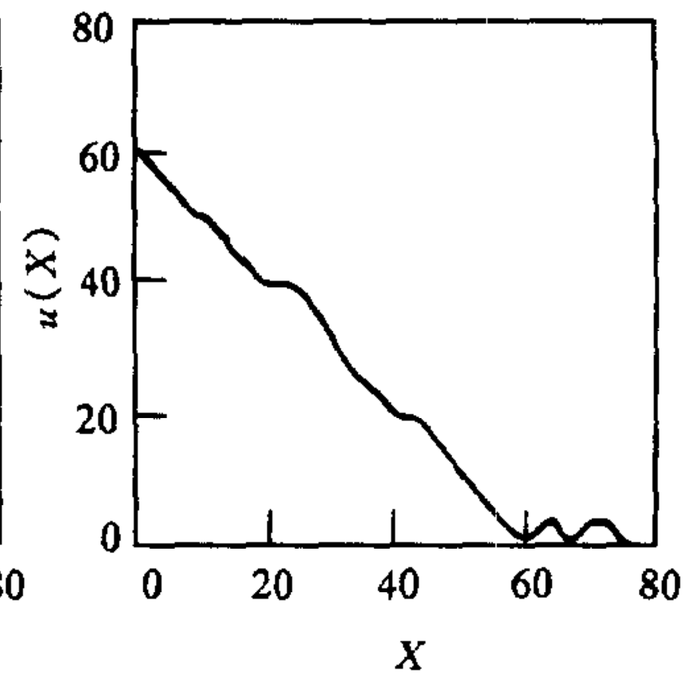

图3

注. 图1 设模型已知时，用值迭代法求解的最优策略 $N=M=50$ $L=2.0$；

图2 设模型未知时，用 Q 学习算法结合新的探索策略求解的控制策略 $N=M=25$ $L=4.0$；

图3 设模型未知时，用 Q 学习算法结合 Boltzmann 分布探索策略求解的控制策略 $N=M=25$ $L=4.0$.

比较图1、图2可知，Q 学习算法所求得的控制策略和在模型已知的情况下用值迭代法求得问题的最优策略非常逼近，均为(6)式所示的 S^* 策略，S^* 约为60.0. 而且仿真表明通过 Q 学习的继续迭代，图2所示的控制策略可以收敛到最优策略. 这一结果说明了在系统模型未知的情况下，通过离散化问题的状态和决策空间，可以用 Q 学习算法来求解某些有连续状态和决策空间的马氏决策问题的最优策略. 同时通过比较图2、图3可知，新的探索策略和 Boltzmann 分布探索相比，不仅提高了学习速度而且求解的控制策略明显更逼近最优解以至能收敛到最优解.

5 结语

本文讨论了 Q 学习算法及相关的探索技术，提出了一种新的探索策略并将该策略和 Q 学习算法有效结合来求解一类典型的有连续状态和决策空间的库存控制问题. 仿真表明：该方法所求解的控制策略和用值迭代法在模型已知的情况下所求得的最优策略非常逼近，从而证实了 Q 学习算法在一些系统模型未知的工程控制问题中的应用潜力.

参 考 文 献

1 Watkins C J C H. Learning from delayed rewards. Ph. D. Dissertation, King's College, UK, 1989

2 Barto A G, Bradtke S J, Singh S P. Learning to act using real-time dynamic programming. *Artificial Intelligence*, 1995, **72**: 81－135

3 Lin L J. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, 1992, **8**: 293－321

4 Berstekas D P. Dynamic Programming and Stochastic Control. New York: Academic Press, 1976

5 Thrun S B. The role of exploration in learning control. Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches. NY: Vanostrand Reinhold, 1992

6 Sutton R S, Barto A G, Williams R J. Reinforcement learning is direct adaptive optimal control. *IEEE Control Sys-*

tem Magazine, 1992, **12**(2):19－22

7 Peng J. Efficient dynamic programming-based learning for control. Ph. D. thesis, Northeastern University, USA, 1993

蒋国飞　1971年生,1993年毕业于北京理工大学自动控制系.现为北京理工大学博士生.主要研究方向是最优控制和智能控制等.

吴沧浦　简介见本刊第20卷第6期.

中国自动化学会1999年一般专题学术活动计划

项目名称	主要内容	时间	人数	地点	联系人
第4届全国农业知识工程学术会议	农业知识工程专业委员会例行学术年会	5月	100	郑州	熊范纶　合肥市1130信箱 邮编　230031
全国机电一体化学术交流会	机电一体化技术的发展、应用与前景展望	5月	100	上海	王志新　上海交通大学机械工程学院　200030
全国神经信息处理与医学工程讨论会	专题研讨	5月	50	湖南	陆惠民　北京朝阳区大屯路15号中科院生物物理所 100101
办公自动化产品和网络技术推广交流	专题研讨	2季度	100	沈阳	李春山　沈阳市东北大学软件中心　邮编:110006
工程设计经验交流会	各大设计院之间的工作经验介绍	2季度	100	北京	朱蕴珍　北京宣武区建工北里1区3号楼1203室 邮编　100053
高压变频器研讨会	交流、介绍国内引进的几家外国公司的高压变频器工作原理、安装、调试及日常维护中的各种技术问题	2季度	80	上海	阮毅　上海大学自动化学院 邮编　200072
全国第二届 Java 技术及应用学术交流会	国内外 Java 技术及应用发展综述,Java 技术软件产品开发及其应用,Java 技术在行业企业中的应用及案例分析等	6月	100	北京	孙凤云　贾志梅　北京海淀区清华东路25号 邮编　100083
第14届青年学术年会	借国际自控联(IFAC)第14届世界大会在北京召开之际,举办青年学者参加的学术交流会,圆桌讲讨论会并组织到国际会议中心的活动	7月3日至6日	100	北京	李小坚　孙力 北京石景山北方工业大学自动化系 100043

(下转263页)