第19卷 第1期 重庆建筑大学学报 Vol.19 No.1
1997年2月 Journal of Chongqiog Jianzhu University Feb.1997

研究简报

马尔科夫质量控制模型

杨春巍
(重庆建筑大学基础科学系 四川重庆 630045)

摘 要 研究了马尔科夫质量控制模型及求解的途径。

关键词 马尔科夫决策规划，马尔科夫质量控制，模型，最优策略
中图法分类号 O221.5

1 MDP 模型

马尔科夫决策规划(Markovian Decision Programming 简称 MDP)是研究动态随机系统的最优化问题。所研究的系统是能够连续、周期地进行观察。在观察时刻，决策者根据观察到的状态，从研究的可付诸实施的决策集中选定一个决策，通过实施，则产生两种结果：

(1) 确定了系统状态概率的转移规律，是具马尔科夫性(无后效性，即所选决策实施的结果与此时刻以前的历史无关)。

(2) 所选定决策实施，将获得一定的经济效益，也具有马尔科夫性。

动态随机系统发展的不同途径将获得不同的经济效益。要求在各个时刻先取决策，使系统处于最优运行状态，即选取最优策略。

假定在时刻 $t=0, 1, 2, \cdots\cdots$ 观察系统，则离散时间 MDP 是一个五元组 $\{S, (A(i), i\in S), q, \Gamma, V\}$ 所构成。

1) S：系统的"状态集"，为一非空集，S 的元素称为状态 $S=S(s, i, j, \cdots\cdots)$。一般，$S$ 为一可列集。

2) $A(i)i\in S$；状态 i 可用的"决策"。一般，$A(i)$为一可列集。

3) q：系统的转移概率，是一族时间上齐次的马尔科夫转移律。每逢系统处于状态 i，选取决策 $a\in A(i)$，则不管系统的历史如何，下次转移到状态 j 的概率为 $q(j\mid j, a)$ 记为 $q_{ij}(a)$。

4) Γ：系统的报酬函数，Γ 是定义在 $\Gamma=\{(i, a): a\in A(i), i\in S\}$ 上的有界单值实函数。每当系统处于状态 i，选取决策 a，则可获得一个报酬 $\Gamma=(i, a)$，它与历史无关。

5) V：系统的目标函数，V 是定义是 $S\cdot\Pi$上的单值实函数，Π是全体策略所成之集。

$\forall t\geqslant 0$ 令

$$h:=\{i_0, a_0, i_1, a_1, \cdots\cdots, i_t, a_t, a_n\in A(i_n), i_n\in S\} \quad n=0,1,2,\cdots\cdots,t$$

收稿日期：1995-05-26
杨春巍，男，1940年生，副教授

称为系统直到时刻 t 的一个"历史"，这样历史的全体构成 H_t，称系统直到时刻 t 的历史集。

$\forall\ h_{t-1}\in H_{t-1}$

$q(j\mid h_{t-1}, i_t, a_t)=q(j\mid i_t, a_t)\quad j\in S\quad t=0, 1, 2, \cdots\cdots$

表示转移概率与历史无关，即马尔科夫性。

2　MDP 决策过程

一个策略(policy) π 是一个序列 $\pi=\{\pi_0, \pi_1, \pi_2, \cdots\cdots\}$，其中 $\forall\ t\geqslant 0$, $h_{t-1}\in H_{t-1}$, $i_t\in S$，$\pi_t(\cdot\mid h_{t-1}, i_t)$ 是 $A(i_t)$ 上的一个概率分布。全体策略集记作 Π。$\forall\ h_{t-1}\in H_{t-1}$, $i_t\in S$ 总存在一个 $a_t\in A(i_t)$，使得 $\pi_t(a_t\mid h_{t-1}, i_t)=1t\geqslant 0$ 则称为一个"决策性策略"，全体决定性策略之集记为 Π^d。

又若一个策略 $\pi=\{\pi_0, \pi_1, \pi_2, \cdots\cdots\}\ \forall\ t\geqslant 0$ 它的 π_t 依赖于时刻 t 所处的状态 i，即 $\pi_t(\cdot\mid h_{t-1}, i_t)\equiv\pi_t(\cdot\mid i_t)$ 则称为"随机马尔科夫策略"，它的全体组成的集称为"随机马尔科夫策略类"，记为 Π_m。一个随机马尔科夫策略 $\pi=\{\pi_0, \pi_1, \pi_2, \cdots\cdots\}$，如果它的每个 π_t 均是一个退化概率分布，则称为"马尔科夫策略"，全体马尔科夫策略所成之集称为"马尔科夫策略类"，记为 Π_m^d。

定义在 S 上的映象 f，映 i 入 $A(i)$，即 $f(i)\in A(i)$, $i\in S$，则称 f 为一个"决策函数"，全体决策函数所成之集记为 F，因而一个策略 $\pi=\{\pi_0, \pi_1, \pi_2, \cdots\cdots\}$ 必存在一串 $f_n\in F$，使得 $\pi\equiv\{f_0, f_1, f_2, \cdots\cdots\}$。我们的成果是在 F 为有限的情况下得到的。

若用 $\bar{y}_t$, Δt 分别表示时刻 t 系统所处的状态和选取的决策，则不同的策略 π，相应的 $\bar{y}_t$, Δt 会有区别，给定初始状态 $i\in S$，选用的策略 π 与系统状态的运动规律 q 一起交互作用决定系统的发展规律，因此 $\bar{y}_t$, Δt 是随机变量序列 $L(\pi)=\{\bar{y}_0, \Delta_0; \bar{y}_1, \Delta_1 \cdots\cdots \bar{y}_t, \Delta_t; \cdots\cdots\}$ 称为 MDP 过程。

给定 MDP 的前三元组 $\{S, (A(i), i\in S), q\}$，一个策略 $\pi\in\Pi$ 及一个初始分布确定了一个 MDP 决策过程。

3　MDP 报酬过程

对每个 MDP 决策过程附以报酬结构，即对每个 MDP 过程定义一个泛函过程，由于 Γ 为有界单值实函数，由 $\pi\in\Pi$ 产生的 MDP 过程 $L(\pi)=\{\bar{y}_0, \Delta_0; \bar{y}_1, \Delta_1; \cdots\cdots\}$。定义 $R_t\equiv R_t(\pi)=\Gamma(\bar{y}_t, \Delta t)$, $t\geqslant 0$。由于 S, $A(i)$ $(i\in S)$ 均为可列集，R_t 仍为一随机变量。即 $\{R_t, t\geqslant 0\}$ 为一随机变量序列。称为"由 π 产生的报酬过程"，简称为"报酬过程"。

由于 Γ 有界，存在期望值 $E_\pi\{R_t\mid \bar{y}_0=i\}=\sum_{j, a} P_\pi\{\bar{y}_t=j, \Delta_t=a\mid \bar{y}_0=i\}\Gamma(j, a)$

$\forall\ N\ (0, 1, 2, \cdots\cdots)$ 令

$$V_N(\pi, i)=E_\pi\left\{\sum_{t=0}^{N} R_t\mid \bar{y}_0=i\right\}=\sum_{t=0}^{N} E_t\{R_t\quad \bar{y}_0=i\}$$

$$=\sum_{t=0}^{N}\sum_{j,a}P_\pi\{\bar{y}_t=j,\Delta_t=a\mid\bar{y}_0=i\}\Gamma(j,a) \tag{1}$$

即，$V_N(\pi,i)$表示用策略 π 在 $t=0$ 时刻从状态 i 出发，系统直到时刻 N 的期望总报酬。$V_N(\pi\cdot)$称为"N 阶段目标函数"或统称为有限阶段目标函数"。

若 $\pi\equiv f^\infty$，即为一平衡策略，则(1)化为 $V_N(f,i)=\sum_{t=0}^{N}\sum_{j}q^t[j\mid i,f(i)]\Gamma(j,f(j))$。其中$q^t[j\cdot t,f(i)]$表示用策略 f 时，$t=0$ 从状态 i 出发 t 时刻转移到 j 的概率。

4 MDP 折扣目标函数

若用长期期望总报酬$\sum_{t=0}^{\infty}E_\pi\{R_t\mid\bar{y}_0=i\}$作目标函数，需要对 Γ 作进一步限制，以保证和式收敛。为此，引进一个折扣因子 $\beta(0<\beta\leqslant 1)$对未来的报酬函数打折扣，当 $t\ (>0)$时刻的单位报酬值为初始时刻 $t=0$ 的 β^t 倍，即 令

$$V_\beta(\pi,i)=E_\pi\left\{\sum_{t=0}^{\infty}\beta^t\Gamma_t\mid\bar{y}_0=i\right\}=\sum_{t=0}^{\infty}\beta^t E_\pi\{\Gamma_t\mid\bar{y}_0=i\}$$

$$=\sum_{t=0}^{\infty}\sum_{j,a}\beta^t P_\pi\{\bar{y}_t=j,\Delta_t=a\mid\bar{y}_0=i\}\Gamma(j,a)$$

表示当用策略 π 时、系统从 $t=0$ 时于状态 i 出发的条件下，系统在无限阶段上的期望折扣总报酬 $V_\beta(\pi\cdot)$(或 $V_\beta(\pi)$)称为折扣目标函数，此时决策函数集折扣模型为$\{S,(A(i),i(\in S),a,\Gamma,V_\beta\}$。即

$$V_\beta(\pi,i)=\sum_{t=0}^{\infty}\beta^t\sum_{j\in S}\sum_{a\in A(i)}\{\bar{y}_t=j,\Delta_t=a\mid\bar{y}_0=i\}\Gamma(j,a)$$

$$\pi\in\Pi,\quad i\in S\quad\cdots\cdots \tag{3}$$

若 $\exists\pi^*\in\Pi,\ \forall\pi\in\Pi,\ i\in S$ 均有

$$V_\beta(\pi^*,i)\geqslant V_\beta(\pi,i)$$

则称 π^* 为关于折和目标为最优的，或称 π^* 为 β 最优的，π^*简称为最优策略。π^*是关于初始状态 i 同时达到最优。

若 $\pi\equiv f$，则 π 为一平稳策略，则(3)成为

$$V_\beta(f,i)=\sum_{t=0}^{\infty}\beta^t\sum_{j}q^t(j\mid i,f(i))\Gamma(j,f(i))\qquad i\in S$$

5 马尔科夫质量控制模型

MDP 由于一些具体问题的特殊结构，使得求最优策略变得容易，因而更具有实用价值。马尔科夫质量控制问题就具有这个特性。

考察一台机器的工作情况、在生产过程中机器所处的两种状态：机器是好的(记作 0)；或

是坏的(记作 1)。

若机器在当天工作开始时是好的。则在第二天开始时仍是好的概率为 $1-\Gamma(0<\Gamma<1)$；若机器已是坏的，在重新调整之前一直保持在坏的状态。如机器是好的，将无费用；如机器是坏的，不加调整则每天损失 C 元，如加以调整(假定能即时完成，比如用一台好机器替换)，则花费 R 元，机器调整为好的。因而有：

1) 系统状态集　$S=\{0, 1\}$。

2) 决策 $A(i)$　$i\in S$，此处为

$A(0)=\{1\}$，　$A(i)=\{1, 2\}$　其中"1"表示不调整，"2"表示调整。

3) 系统转移概率 $q(j\mid i, a)$　$a\in A(i)$。

具体为　　$q(0\mid 0, 1)=1-\Gamma$；

$q(1\mid 0, 1)=\Gamma$；　$q(1\mid 1, 1)=1$；

$q(0\mid 1, 2)=1-\Gamma$；　$q(1\mid 1, 2)=\Gamma$。

4) 系统的报酬函数　$\Gamma=\{(i, a)\mid a\in A(i)，i\in S\}$

则有　　$\Gamma(0, 1)=0$；

$\Gamma=(1, 1)=C$；　$\Gamma=(1, 2)=R$

5) 系统的目标函数 V，是寻求一个质量控制策略，即最小期望折扣总费用或最小期望平均费用。

这是一个马尔科夫决策规划问题。

6　马尔科夫质量控制的特殊结构

由于机器所处的状态是不可直接观察的，只能由每天最初开动机器生产出的产品质量来判断，即由后验概率来判断机器所处的状态。设机器处于坏状态的后验概率为 P，则一台好机器的后验概率 $P=0$，坏机器的后验概率 $P=1$。若在当天开始时机器是坏的后验概率为 P，则在第二天开始时机器是坏的后验概率为 $\Gamma+(1-\Gamma)P=TP$，其中 P 可以通过历史资料来获得(若无历史资料，只有用主观概率决定)。

在当天开始时，我们估计机器是坏的后验概率是 P，然后决定：是重新调整机器到好状态，要花费用 R；还是继续生产一天。一般的情况，若当天开始时机器是坏的后验概率为 P，且不调整，一天的费用为 $L(P)$，可理解为生产废品的损失。

将后验概率 P 作为状态，则得到另一个模型。其中：

1) 状态集：　$S=\{P，P\in[0, 1]\}$

2) 决策　　$A(P)=\{1, 2\}$。其中"1"表示不同调整，"2"表示调整。

3) 系统转移概率

$$q(P'\mid P, 1)=\begin{cases}1 & \text{当 } P'=TP；\\ 0 & \text{当 } P'\neq TP。\end{cases}$$

$$q(P'\mid P, 2)=\begin{cases}1-\Gamma & \text{当 } P'=0；\\ \Gamma & \text{当 } P'=1；\\ 0 & \text{其 他。}\end{cases}$$

4) 系统报酬函数

$$\Gamma=(P,1)=L(P)\quad;\qquad \Gamma=(P,2)=R\quad 。$$

5) 系统的目标函数 V，仍是寻求一个质量控制策略，即最小期望折扣总费用，或最小期望平均费用。

这是马尔科夫质量控制模型。

7 马尔科夫质量控制的求解

若当天开始时，机器为坏的后验概率为 P；当天开始时机器是好的，下一天开始时仍是好的后验概率是 $1-\Gamma$；当天开始时机器是坏的，下一天开始时仍是坏的后验概率就是

$$TP=\Gamma+(1-\Gamma)P\quad 。$$

若当天开始时机器是坏的，下一天开始时机器为坏的概率为 P；则当天开始时机器是好的，下一天开始时仍是好的后验概率为

$$T^2P=(1-P)(1-\Gamma)^2\quad 。$$

这样，就有当天开始时机器是坏的后验概率为 P，则下几天开始时机器为坏的后验概率为

$$T^nP=1-(1-P)(1-\Gamma)^n\qquad n=1,2,3,\cdots\cdots$$

由此可知状态集 S 仅为可数集，这样的问题我们已有专文论述了求它的最优策略的方法（参考文献[2]）。

可是利用问题的特殊性，可以得到更好的结果。

令

$$\Delta L(T^{t+1},0)=L(T^{t-1},0)-L(T^{t},0)$$

若 $L(P)\geqslant 0$ 则存在一个最优策略为：

$$V_\beta(\pi,i)=L(P)=G_\beta(t(\beta))$$

假定现在的后验概率 $P>P(\beta)$，则重新调整机器；否则（即 $P\leqslant P(\beta)$ 就维持生产，其中，当

$$\sum_{t=1}^{\infty}\Big(\sum_{k=0}^{t}\beta^k\Big)\Delta(T^{t+1},0)\leqslant R$$

时，$P(\beta)=1$，否则 $P(\beta)<1$，它是

$$L(P)=G_\beta(t(\beta))$$

的唯一解。此处

$$G_\beta(t(\beta))=\frac{\Big(\sum_{t=0}^{t}\beta^t L(T^t,0)+R\Big)}{\sum_{t=0}^{t}\beta^t}$$

$$t(\beta)=\min\Big\{t:\sum_{t=0}^{t}\Big(\sum_{k=0}^{k}\beta^t\Big)\Delta L(T^{t+1},0)\geqslant R\Big\}$$

其中 t 为整数。

当 $\beta \uparrow 1$ 时，　$P(\beta) \downarrow P(1)$，且 $t(\beta) \downarrow t(1)$ ；

而当 $\beta < 1$ 时，为折扣模型的最优策略；　当 $\beta = 1$ 时，为平均目标的最优策略。

这些都化为典型的 MDP 模型，很好的解决了问题。

参 考 文 献

1 杨春巍．F 有限折扣模型的策略迭代法．重庆建筑工程学院学报，1988(2)

2 A. T. Bharucha－Reid. 马尔科夫过程论及其应用．上海．上海科学技术出版社，1979

3 董泽清．马尔科夫决策规划．中国科学院应用数学研究所，1985

A model of Markovian qualty control

Yang　Chunwei

(Dept. of Natural Science, Chongqing Jianzhu University, Chongqing, Sichuan 630045)

Abstract: Markovian qualty confrol is an application of Markovian decision programming. It has its special structure. This paper studies a model of Markovian quality control and the approach for solving these prohlems

Key Words: Markoian decision programming, Markonvian qualty control model, optimal policy.

(编辑·袁　江)

科研成果

新型磷渣硅酸盐水泥

内容简介及技术水平：

新型磷渣硅酸盐水泥技术创造性地选用了价廉高效的多功能天然矿物或人造复合矿物代替石膏生产水泥，使磷渣水泥的性能得到明显改善。具有国际先进水平，其经济效益、社会效益和环境效益十分显著。