第 4 章　Markov 决策规划

4.1　引　言

所谓决策, 是指在若干个可行的行动方案中按照某种准则选出一个方案. 其中, 有一类多阶段决策问题称为序贯决策, 即在系统的运行过程中, 它不是作一次决策就结束, 而是在一系列观察的时刻点上都要作出决策. 如一家商店各种商品每月的进货量; 一台机器定期的维修; 一家工厂每月的生产计划等. 在每个观察时刻点上, 决策者首先根据观察所得的系统状态, 从其所有备选方案中选择一个方案 (即作出决策) 执行, 其结果是: (1) 将获得一定效益; (2) 能确定以后系统状态发展的概率规律. 然后, 再观察下一时刻点上系统出现的状态, 据此再次作新的决策, 如此一步一步地进行下去. 如果在序贯决策过程中, 系统状态的转移服从已知的概率规律且与系统以前的发展历史无关, 即具有无后效性 (或 Markov 性), 称此类序贯决策问题的数学模型为 Markov 决策规划 (以下简记为 MDP).

Markov 决策规划 是解决随机性序贯决策问题的重要分支学科. 它可以应用于许多领域, 是解决随机动态最优化问题的重要工具. 如排队系统的最优运行控制; 随机库存系统的最优订货策略; 设备的最优更换维修策略; 水库的优化调度等均可以化为一定的 MDP 来解决. 可以说, 凡是以 Markov 过程作为数学模型的问题, 只要能够引入 “行动”与 “报酬”结构, 均可以应用 Markov 决策规划. 本章中对此作一简单的介绍, 有关详细的讨论, 请参见文献 [5].

4.2　Markov 决策规划的数学描述

一般地说, Markov 决策规划可由一个五重组 $\{S, (A(i), i\in S), P, r, V\}$ 来描述. 下面, 先介绍一个典型的机器维修问题.

4.2.1 机器维修问题

【例 4.1】(机器维修最优策略问题) 设等周期 (如一天) 地考察一台运行的机器, 在每周期初始时刻观察它的运行情况. 每次观察时, 机器可处于以下两个状态之一: 正常运行 (记做 $i=1$) 或出了故障 (记做 $i=2$). 在任一周期, 若机器正常运行可得收益 10 元, 到下一周期初, 仍处于正常运行的概率为 0.7, 处于出故障状态的概率为 0.3. 处于正常运行状态时, 可用的行动只有一个, 即继续生产 (记做 a_1). 若处于故障状态 2, 则有两个行动可供选择: 快修 (记做 a_2) 和常规修理 (记做 a_3). 在快修时需付费用 5 元, (即收益为 -5 元), 而该时段能修复为正常运行状态的概率为 0.6; 在常规修理时需付费用 2 元, 且在该时段能修复的概率为 0.4. 问题是: 在各个周期初, 根据观察到的系统实际所处的状态, 如何选取行动, 才能使整个考察期内的某种期望收益达到最大.

解 容易看出, 机器可能处于两种状态, 记状态空间 $S=\{1,2\}$, 每种状态下可采用行动方案有 $A(1)=\{a_1\}, A(2)=\{a_1,a_2\}$. 用 $P(j\,|i,a)$ 表示在时刻 t 观察到系统状态为 i, 选用方案 a, 于 $(t+1)$ 时刻转移到状态 j 的概率. $r(i,a)$ 表示在时刻 t 观察到系统状态为 i, 选用方案 a 时所获得的收益. 此处 $P(j\,|i,a)$ 与 $r(i,a)$ 都与 t 时刻以前系统的历史无关. 将 $P(j\,|i,a)$ 和 $r(i,a)$ 的数值列表, 见表 4.1.

表 4.1 转移概率与报酬

状态 (i)	行动 (a)	转移概率 $P(j\|i,a)$		报酬 (元)
		$j=1$	$j=2$	$r(i,a)$
1	a_1	0.7	0.3	10
2	a_2	0.6	0.4	-5
	a_3	0.4	0.6	-2

决策规则 (或称方案选择规则) f 表示如下的映射: 当观察到系统状态为 1 时, 选择方案 a_1; 当观察到系统状态为 2 时, 选择方案 a_2. 即 $f(1)=a_1, f(2)=a_2$. 类似地, 可令决策规则 g 表示如下映射: $g(1)=a_1, g(2)=a_3$.

当 $t=0$ 时, 根据选择规则从 f,g 中选用一个决策, 记为 f_0 (相应得出选用的方案), 从状态 i 出发获得收益 $r(i,f_0(i))$; 当 $t=1$ 时, 机器转移到状态 j 的概率为 $P(j\,|i,f_0(i))$ $(i,j=1,2)$, 同样从 f,g 中选用一个决策, 记为 f_1. 由于状态转移是随机的, 因而获得的收益也是随机的, 其期望收益为 $\sum\limits_{j=1}^{2} P(j|i,f_0(i))r(j,f_1(j))$. 当 $t=2$ 时, 机器转移到状态 k 的概率为 $P(k|j,f_1(j))$ $(k,j=1,2)$. 再从 f,g 中选用一个决策, 记为 f_2. 依次下去, 得一

决策序列 $(f_0, f_1, f_2, f_3, \cdots)$ (相应可得到方案序列), 将其记为 π, 称为策略. 由于收益是从 $t=0$ 时开始计算的, 考虑到经济上利率的影响, 则在 t 时段的单位收益可折合成初始时刻 $t=0$ 时的值 β^t, 其中 $\beta=\dfrac{1}{1+a}$ $(a>0)$. 因此, $t=0$ 时从状态 i 出发, 长期的期望折扣总收益为

$$\begin{aligned} V_\beta(\pi,i) =& r(i,f_0(i))+\beta\sum_{j=1}^{2}P(j|i,f_0(i))r(j,f_1(j)) \\ &+\beta^2\sum_{k=1}^{2}\sum_{j=1}^{2}P(j|i,f_0(i))P(k|j,f_1(j))r(k,f_2(k))+\cdots \quad (i=1,2) \end{aligned}$$

$V_\beta(\pi,i)$ 就是衡量本问题策略优劣的准则. 当系统的状态转移律已知时, 它显然就是初始状态 i 和策略 π 的函数, 本问题就是寻求这样的方案序列 π, 使 $V_\beta(\pi,i)$ 获得最大值.

4.2.2 MDP 的数学描述

一个离散时间的 MDP 模型是由 $\{S,(A(i),i\in S),P,r,V\}$ 五个部分组成的.

状态空间记为 S, 它是被观察的系统所有可能状态的集合, 常用小写字母 $i,j,k,\cdots$ 表示.

行动集记为 $A(i)$, 它是系统处于状态 i 时, 决策者可采用的行动的集合. 常用小写字母 $a,b,c,\cdots$ 表示.

P 是系统状态的齐次的 Markov 转移律, $P(j\ |i,a)$ 表示任一时刻 t $(t=0,1,2,\cdots)$ 系统处于状态 i, 选用方案 $a\in A(i)$ 后, 在 $(t+1)$ 时刻转移到状态 j 的概率. 它与系统在 t 时刻之前的历史及时刻 t 均无关, 且满足 $P(j|i,a)\geqslant 0$, $\sum\limits_{j\in S}P(j|i,a)=1$. 其中 $a\in A(i), i,j\in S$.

收益函数 r 是定义在 $\Gamma=\{(i,a)|a\in A(i),i\in S\}$ 上的单值实函数, 记为 $r(i,a)$. 它表示在任一时刻 t, 系统处于状态 i, 选用方案 $a\in A(i)$ 时所获得的收益 (负的收益就是费用, 支出).

目标 V 是定义在 $\Pi\times S$ 上的单值实函数, 对任意给定的策略 $\pi\in\Pi, i\in S, V(\pi,i)$ 表示 $t=0$ 时从状态 i 出发, 用策略 π 所获得的目标值. 它是衡量不同策略优劣的标准.

当给定一组 $\{S,(A(i),i\in S),P,r,V\}$ 时, 就认为给定了一个具体的 MDP 模型. 在本书中, 为简单起见, 假定状态空间 S 及行动集 $A(i)$ $(i\in S)$ 均为有限集.

4.2.3 策略类与目标函数

从状态空间 S 到备选方案集 $A(i)$ 的映射 f 叫做方案选择规则 (或称为决策规则、决策函数). 全体决策规则所成之集记为 F .

【定义 4.1】 设 $f_t \in F, t \in \{0,1,2,\cdots\}, f_t$ 是在时刻 t 选择行动的规则, 令决策规则序列 $\pi = (f_0, f_1, f_2, \cdots)$, 其中. 如果它不依赖于时刻 t 以前系统的历史, 只要知道时刻 t 及该时刻系统的状态 i, 按照 π 选择的方案 $f_t(i)$ 就惟一决定了, 称 π 为 Markov 策略. 全体 Markov 策略所成之集记为 Π_m^d, 称为 Markov 策略类.

【定义 4.2】 在 Markov 策略中, 于 t 时刻选择方案的规则 (记为 π_t) 具有随机性, 即在时刻 t, 系统处于 i_t 时选用方案 a 的概率为 $\pi_t(a|i_t)$, 且 $\pi_t(a|i_t) \geqslant 0, \sum\limits_{a \in A(i)} \pi_t(a|i_t) = 1$, 则称这种 $\pi = (\pi_0, \pi_1, \pi_2, \cdots)$ 为随机 Markov 策略. 全体随机 Markov 策略所成之集记做 Π_m, 称为随机 Markov 策略类.

若在 t 时刻选择方案的规则 π 不仅是随机的, 而且还依赖于 t 时刻以前系统的历史, 则称 $\pi = (\pi_0, \pi_1, \pi_2, \cdots)$ 为一般策略, 简称策略. 全体策略所成之集记做 Π, 称为策略空间.

容易看出, 各策略类之间的关系 $\Pi_m^d \subset \Pi_m \subset \Pi$.

在 MDP 中常用的目标有以下三种:

(1) 有限阶段目标 $V_N(\pi)$

对给定的自然数 N, 选用策略 $\pi \in \Pi$ 及 $i \in S$, 令

$$V_N(\pi, i) = \sum_{t=0}^{N} \sum_{\substack{a \in A(i) \\ j \in S}} P_\pi\{Y_t = j, \Delta_t = a | Y_0 = i\} r(j, a) \tag{4.1}$$

其中 Y_t, Δ_t 分别表示在时刻 (或阶段) t 所观察到的系统状态和决策者所选用的行动. $P_\pi\{Y_t = j, \Delta_t = a | Y_0 = i\}$ 表示用策略 $\pi \in \Pi$, 在 $t = 0$ 从状态 i 出发, 选用行动 $a \in A(i)$ 于阶段 t 转移到状态 j 的条件概率; $V_N(\pi, i)$ 表示采用策略 π, 在 $t = 0$ 时从状态 i 出发直到时刻 N 所获得的期望总收益. 设 $V_N(\pi)$ 表示一列向量, 其第 i 个分量为 $V_N(\pi, i)(i \in S)$, 称 $V_N(\pi)$ 为 N 阶段目标 (准则, 指标). 有时为了方便, 不指明阶段数 N, 而统称为"有限阶段目标". 以有限阶段作为目标的 MDP 称为有限阶段模型.

综合上述定义, 当五重组 $\{S, (A(i), i \in S), P, r, V_N\}$ 给定了以后, 就认为定义了一个有限阶段 MDP 模型.

(2) 折扣目标 $V_\beta(\pi)$

在前面介绍的内容中, 没有考虑决策者对时间的偏好. 实际上, 在许多经济问题中, 决策人对目前的收益 (或损失) 要比对未来的收益 (或损失) 看得重. 于是, 在把未来的收益折算成现在的收益时, 需要乘上一个折扣率, 也即贴现率, 记做 β $(0<\beta<1)$. 如把未来的收入折合成 $t=0$ 时的收入, 即第 t 周期的单位收入折合成 $t=0$ 时的值为 β^t. 对有限阶段的折扣目标 $V_{N,\beta}(\pi,i)$, 有

$$V_{N,\beta}(\pi,i)=\sum_{t=0}^{N}\beta^t\sum_{\substack{a\in A(i)\\ j\in S}}P_\pi\{Y_t=j,\Delta_t=a|Y_0=i\}r(j,a) \tag{4.2}$$

$V_{N,\beta}(\pi,i)$, 表示选用策略 π, 在 $t=0$ 时从状态 i 出发至时刻 N 所获得的折扣期望总收益. 若考虑无穷阶段 (或长期) 的期望总收益最大 (或期望总费用最小), 则有

$$V_\beta(\pi,i)=\sum_{t=0}^{\infty}\beta^t\sum_{\substack{a\in A(i)\\ j\in S}}P_\pi\{Y_t=j,\Delta_t=a|Y_0=i\}r(j,a) \tag{4.3}$$

$V_\beta(\pi,i)$ 表示用策略 π, 在 $t=0$ 时系统从状态 i 出发的条件下无限阶段的折扣期望总收益. 用 $V_\beta(\pi)$ 表示一列向量, 其第 i 个分量为 $V_\beta(\pi,i)$. 称 $V_\beta(\pi)$ 为 "折扣准则 (目标)".

五重组 $\{S,(A(i),i\in S),P,r,V_\beta\}$ 定义了一个折扣 MDP 模型.

(3) 平均目标 $\overline{V}(\pi)$

对任给的 $\pi\in\Pi, i\in S$, 令

$$\overline{V}_N(\pi,i)=\frac{V_N(\pi,i)}{N+1}$$

其中 $V_N(\pi,i)$ 为有限阶段目标, 则 $\overline{V}_N(\pi,i)$ 表示采用策略 π 在 $t=0$ 时从状态 i 出发, 前 $N+1$ 阶段的平均期望收益.

如果考虑使无限阶段的平均期望收益达到最大, 令

$$\begin{aligned}\overline{V}(\pi,i)&=\lim_{N\to\infty}\inf\frac{V_N(\pi,i)}{N+1}\\&=\lim_{N\to\infty}\inf\frac{1}{N+1}\sum_{t=0}^{N}\sum_{\substack{a\in A(i)\\ j\in S}}P_\pi\{Y_t=j,\Delta_t=a|Y_0=i\}r(j,a)\end{aligned}$$

$\overline{V}(\pi,i)$ 表示用策略 π, 在 $t=0$ 时从状态 i 出发长期每单位时间的平均收益. 令 $\overline{V}(\pi)$ 为一列向量, 其第 i 个分量为 $\overline{V}(\pi,i)$, 则称 $\overline{V}(\pi)$ 为长期每单位时间平均期望目标, 简称平均目标, 以平均目标为准则的 MDP 称为平均准则 Markov 决策规划, 详细的讨论请参见文献 [5, 26] 等.

与一般的最优化问题一样, MDP 首先要解决 “最优”的概念问题. 因此, 在本节中首先定义最优策略与最优值函数的概念, 然后讨论在上述最优性的定义下, 对于各种模型, 其最优策略是否存在等一些问题, 并给出求解最优策略的有效算法.

4.3 有限阶段模型

由前所述, 有限阶段模型是指在一个 MDP 模型 $\{S,(A(i),i\in S),P,r,V\}$ 中, 目标函数的收益只计算到给定的有限阶段 N. 即对于给定的自然数 N, 选用策略 π 及状态 $i\in S$, 决策者在 $t=0$ 时从状态 i 出发直到时刻 N 所获得的总收益. 可由如下形式来计算:

$$V_N(\pi,i)=\sum_{t=0}^{N}\sum_{\substack{a\in A(j)\\ j\in S}}P_\pi\{Y_t=j,\Delta_t=a|Y_0=i\}r(j,a)$$

下面将介绍有限阶段模型中最优策略的概念, 最优策略的存在性问题及其求解算法.

4.3.1 最优策略及其存在性

【定义 4.3】 *在有限阶段模型中, 若存在一个策略 $\pi^*\in\Pi$, 使得对任何 $\pi\in\Pi, i\in S$ 均有*

$$V_N(\pi^*,i)\geqslant V_N(\pi,i)$$

则称 π^* 关于有限阶段准则在 Π 中是最优的, 简称为最优策略. $V_N(\pi^*)=(V_N(\pi^*,1),V_N(\pi^*,2),\cdots,V_N(\pi^*,l))$ 称为关于有限准则的最优值函数, 简称为最优值函数. 其中 $S=\{1,2,\cdots,l\}$.

这里的最优是关于所有初始出发状态 i 同时达到最优, 这种最优性概念自然是很强的.

可以证明, 在有限阶段模型 $\{S,(A(i),i\in S),P,r,V_N\}$ 中, 设 S 及 $A(i)(i\in S)$ 均为有限集, 则必有一 Markov 策略 $\pi^*=(f_0^*,f_1^*,\cdots,f_N^*),f_t^*\in F$ $(t=$

$0,1,\cdots,N$), 它关于有限阶段准则是最优的. 即对任何 $\pi\in\Pi, i\in S$ 有

$$V_N(\pi^*,i)\geqslant V_N(\pi,i)$$

详细证明请参见文献 [5, 26, 27].

4.3.2 向后归纳法

在确定性动态规划问题求解中, 向后归纳法是寻求最优策略的有效解法, 同样, 也是求解有限阶段 Markov 决策规划问题中最优策略与最优值函数的有效解法.

由文献 [28] 有

【定理 4.1】 在状态集与所有行动集均为有限的有限阶段模型中, 定义函数 $V_*^n(i)$, 使其满足如下等式:

$$\begin{aligned} V_*^n(i) &= \max_{a\in A(i)}\left[r(i,a)+\sum_{j\in S}P(j|i,a)V_*^{n+1}(j)\right] \\ &= r(i,f_n^*(i))+\sum_{j\in S}P(j|i,f_n^*(i))V_*^{n+1}(j) \qquad (4.4) \\ &\quad (i\in S, n=N,N-1,N-2,\cdots,0) \end{aligned}$$

其中 $V_*^{N+1}(j)=0$. 则由上述算式求出的 $V_*^0=(V_*^0(1),V_*^0(2),\cdots,V_*^0(l))$ 即为有限阶段模型的最优值函数, 即对每个 $i\in S$, 均有 $V_*^0(i)=\sup\limits_{\pi\in\Pi}V_N(\pi,i)$; 与此同时求得的决策序列 $\pi^*=(f_0^*,f_1^*,\cdots,f_N^*)$ 即为最优策略, 其中设 $S=\{1,2,\cdots,l\}$.

由于所有的 $A(i)(i\in S)$ 及 $S=\{1,2,\cdots,l\}$ 均为有限集, 故由 (4.4) 式求得的 $f_n^*(i)$ 一定存在, 且达到最大的行动可能多于一个 (此时可任取一个作为 $f_n^*(i)$). 定理 4.1 不仅解决了有限阶段模型求解最优策略的方法问题, 而且还表明, 对任何 n, $V_*^n(i)$ 表示在阶段 n, 从状态 i 出发, 在余下的 $N+1-n$ 阶段的最优期望总报酬, $(f_n^*,f_{n+1}^*,\cdots,f_N^*)$ 也构成从 n 到阶段 N 的最优策略. 这也体现了 Bellman 的 “最优化原理”.

【例 4.2】 试求例 4.1 中当 $N=3$ 时的最优策略与最优值函数.

解 由题意知, 机器只有两个状态, 即 $S=\{1,2\}$, 对应的行动集分别为 $A(1)=\{a_1\}, A(2)=\{a_1,a_2\}$. 故最优值函数的形式为 $V_*^0=(V_*^0(1),V_*^0(2))$, 其中

$V_*^0(1)$ 与 $V_*^0(2)$ 可通过 (4.4) 式分别求解得到. 注意到题设取 $N=3$, 因而根据向后归纳法的求解顺序应为 $V_*^4(i)\to V_*^3(i)\to V_*^2(i)\to V_*^1(i)\to V_*^0(i)$, 其中 $i\in S=\{1,2\}$. 下面分别列出 $n=3,2,1,0$ 时按照 (4.4) 式计算的有关结果.

(1) $n=3$, 有

$$\begin{aligned} V_*^4(1) &= V_*^4(2)=0\\ V_*^3(1) &= \max_{a\in A(1)}\{r(1,a)+\sum_{j\in S}P(j|1,a)V_*^4(j)\}\\ &= \max_{a\in A(1)}\{r(1,a)\}=r(1,a_1)=10 \end{aligned}$$

达到 $V_*^3(1)$ 右边最大的行动为 a_1, 故令 $f_3^*(1)=a_1$;

$$\begin{aligned} V_*^3(2) &= \max_{a\in A(2)}\{r(2,a)+\sum_{j\in S}P(j|2,a)V_*^4(j)\}\\ &= \max_{a\in A(2)}\{r(2,a_2),r(2,a_3)\}=\max\{-5,-2\}=-2 \end{aligned}$$

达到右端最大的行动为 a_3, 故令 $f_3^*(2)=a_3$.

(2) $n=2$, 由 (4.4) 式及上一步计算得到的 $V_*^3(1),V_*^3(2)$, 有

$$\begin{aligned} V_*^2(1) &= \max_{a\in A(1)}\{r(1,a)+\sum_{j\in S}P(j|1,a)V_*^3(j)\}\\ &= r(1,a_1)+0.7\times 10+0.3\times(-20)=16.4 \end{aligned}$$

故令 $f_2^*(1)=a_1$;

$$\begin{aligned} V_*^2(2) &= \max_{a\in A(2)}\{r(2,a)+\sum_{j\in S}P(j|2,a)V_*^3(j)\}\\ &= \max\{r(2,a_2)+0.6\times 10+0.4\times(-2),\\ &\qquad r(2,a_3)+0.4\times 10+0.6\times(-2)\}\\ &= \max\{0.2,0.8\}=0.8 \end{aligned}$$

达到 $V_*^2(2)$ 右端最大的行动为 a_3, 故令 $f_2^*(2)=a_3$.

(3) $n=1$, 由 (4.4) 式及上一步计算得到的 $V_*^2(1),V_*^2(2)$, 有

$$\begin{aligned} V_*^1(1) &= \max_{a\in A(1)}\{r(1,a)+\sum_{j\in S}P(j|1,a)V_*^2(j)\}\\ &= 10+0.7\times 16.4+0.3\times 0.8=21.72 \end{aligned}$$

故令 $f_1^*(1)=a_1$;

$$\begin{aligned}V_*^1(2)&=\max_{a\in A(2)}\{r(2,a)+\sum_{j\in S}P(j|2,a)V_*^2(j)\}\\&=\max\{r(2,a_2)+0.6\times16.4+0.4\times0.8,\\&\qquad r(2,a_3)+0.4\times16.4+0.6\times0.8\}\\&=\max\{5.16,5.04\}=5.16\end{aligned}$$

达到 $V_*^1(2)$ 右端最大的行动为 a_2, 故令 $f_1^*(2)=a_2$.

(4) $n=0$, 由 (4.4) 式及上一步计算得到的 $V_*^1(1),V_*^1(2)$, 有

$$\begin{aligned}V_*^0(1)&=\max_{a\in A(1)}\{r(1,a)+\sum_{j\in S}P(j|1,a)V_*^1(j)\}\\&=10+0.7\times21.72+0.3\times5.16=26.752\end{aligned}$$

故令 $f_0^*(1)=a_1$;

$$\begin{aligned}V_*^0(2)&=\max_{a\in A(2)}\{r(2,a)+\sum_{j\in S}P(j|2,a)V_*^1(j)\}\\&=\max\{-5+0.6\times21.72+0.4\times5.16,-2+0.4\times21.72+0.6\times5.16\}\\&=\max\{10.096,9.784\}=10.096\end{aligned}$$

达到 $V_*^0(2)$ 右端最大的行动为 a_2, 故令 $f_0^*(2)=a_2$.

由定理 4.1 得知最优值函数为

$$V_*^0=(V_*^0(1),V_*^0(2))=(26.752,10.096)=(V_3(\pi^*,1),V_3(\pi^*,2))$$

相应的最优策略为 $\pi^*=(f_0^*,f_1^*,f_2^*,f_3^*)=(f,f,g,g)$, 其中 $f(1)=g(1)=a_1,f(2)=a_2,g(2)=a_3$.

注: 本例中的最优策略不是平稳的, 决策函数 f_2,f_1,f_0 不同. 有限阶段问题的最优策略一般不是平稳策略.

【例 4.3】 假设一设备制造厂承接了某工程中一台关键设备的制造任务. 工程对此设备的质量标准有非常严格的要求. 以该厂现有的技术水准而言, 每台制成的设备能通过质量检验而被接受的概率仅为 0.25. 再因该工程对此设备又有一定的时限要求, 所以厂方决定, 至多安排三个生产周期完成此项任务. 每一生产周期可制造 j $(0\leqslant j\leqslant 3)$ 台设备. 在每一生产周期结束时, 均对已制成的设备进行检验. 只要其中有一台是合格的, 便不再安排新的生产周期. 在每一生

产周期, 只要开工制造这种设备便需一固定的开工费用 C_1. 此外, 生产每台设备的费用为 C_2. 若在第三个生产周期结束时, 厂方仍未能生产出一台合格的设备, 从而无法向工程供货, 则需履行事先签定的合同, 向工程方面支付一笔违约费用 C_3. 试问, 厂方应制定怎样的生产策略, 以使期望总费用最小?

解 此例中生产周期至多为 3, 故取 $N+1=3$, 即 $N=2$. 在每一生产周期结束时厂方只关心是否已制造出合格设备, 取状态空间 $S=\{0,1\}$, 状态 0 表示厂方已制造出合格设备, 状态 1 则表示还未制造出合格设备. 以行动 j $(0 \leqslant j \leqslant 3)$ 表示在一生产周期内生产 j 台设备, 则有如下行动集为 $A(0)=\{0\}, A(1)=\{0,1,2,3\}$. 再以 $r(i,j)$ 表示状态为 i 时采取行动 j 所导致的费用, 则有

$$r(i,j)=\begin{cases} C_1+C_2 j, & i=1, j>0 \\ 0, & \text{其他} \end{cases}$$

最终费用函数用 $R(i)$ 表示, 有

$$R(i)=\begin{cases} 0, & i=0 \\ C_3, & i=1 \end{cases}$$

最后, 根据题意, 若在一生产周期内制造了 j 台设备, 则此 j 台设备均被拒收的概率为 $(\frac{3}{4})^j$. 于是, 转移概率族用如下转移概率矩阵形式:

$$P(0)=\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, P(1)=\begin{bmatrix} 1 & 0 \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}, P(2)=\begin{bmatrix} 1 & 0 \\ \frac{7}{16} & \frac{9}{16} \end{bmatrix}, P(3)=\begin{bmatrix} 1 & 0 \\ \frac{37}{64} & \frac{27}{64} \end{bmatrix}$$

假定 $C_1=10, C_2=5, C_3=64$, 于是, 有

$$r(i.j)=\begin{cases} 10+5j, & i=1, j>0 \\ 0, & \text{其他} \end{cases}, \quad R(i)=\begin{cases} 0, & i=0 \\ 64, & i=1 \end{cases}$$

下面将递推地计算最优值函数并确定相应的最优策略.

首先考虑状态 0. 由于 $A(0)=\{0\}$, 且 $V_*^3(0)=R(0)=0$, 故

$$\begin{aligned} V_*^2(0) &= r(0,0)+\sum_{j\in S} P(j|0,0)V_*^3(j) \\ &= r(0,0)+\sum P(0|0,0)V_*^3(0)=V_*^3(0)=0 \end{aligned}$$

类似可求得, $V_*^1(0)=V_*^0(0)=0$. 于是显然有 $f_0^*(0)=f_1^*(0)=f_2^*(0)=0$.

其次, 考虑状态 1. 作为初始值有 $V_*^3(0)=0, V_*^3(1)=64$. 下面依次递推以求出 $V_*^2(1), V_*^1(1), V_*^0(1)$.

由于 $A(1)=\{0,1,2,3\}$, 有

$$\begin{aligned} V_*^2(1) &= \min_{a\in A(1)}\{r(1,a)+\sum_{j\in S}P(1|1,a)V_*^3(1)\} \\ &= \min\{r(1,0)+1\times 64, r(1,1)+64\times\frac{3}{4}, \\ &\qquad r(1,2)+64\times\frac{9}{16}, r(1,3)+64\times\frac{27}{64}\} \\ &= \min\{0+64, 15+48, 20+36, 25+27\}=52 \end{aligned}$$

故令 $f_2^*=3$.

再由 $V_*^2(0)=0, V_*^2(1)=52$, 可得

$$\begin{aligned} V_*^1(1) &= \min_{a\in A(1)}\{r(1,a)+\sum_{j\in S}P(1|1,a)V_*^2(1)\} \\ &= \min\{0+52\times 1, 15+52\times\frac{3}{4}, 20+52\times\frac{9}{16}, 25+52\times\frac{27}{64}\} \\ &= 25+52\times\frac{27}{64}=46.9375 \end{aligned}$$

故可令 $f_1^*(1)=3$.

最后, 类似可求得 $V_*^0(1)=44.801758, f_0^*(1)=3$.

于是, 本题中三阶段模型的最优值函数为 $V_*^0=(V_*^0(0),V_*^0(1))=(0,44.801758)$, 最优策略为 $\pi^*=(f_0,f_1,f_2)$, 其中 $f_i(0)=0, f_i(1)=3$ $(\forall\ 0\leqslant i\leqslant 2)$.

这表明厂方采取这样的策略: 在每一生产周期结束时, 只要还未制造出合格的设备, 便在下一周期生产三台设备; 若至少已制造出一台合格设备, 便终止生产. 此处, 在第三个生产周期结束时, 生产也自动停止. 显然, 如厂方最初有合格设备的库存, 则立即交货, 从而费用为 0; 否则, 在采取上述生产策略后, 可使期望总费用达最小, 即 44.801758.

直观上不难看出, 因最终惩罚费用相对地要比固定的与可变的生产费用水平大许多, 厂方采取上述策略是很自然的, 此处可以想像, 一旦费用结构改变, 最优策略也将相应地有所改变.

用有限阶段模型的向后归纳法来求解 Markov 决策规划问题虽然方法较简单, 但前提是要确切知道该序贯决策问题将在某有限时段内结束. 然而, 很多实际情况是人们往往无法确定该系统在什么时候结束, 即使知道它在有限时间结束, 但阶段数 $N+1$ 很大, 导致了较大的计算量, 因而还需考虑别的模型.

4.4 折扣模型

由前所述, 折扣模型是指在一个 MDP 模型中, 其目标函数采用如下形式:

$$V_\beta(\pi,i)=\sum_{t=0}^{\infty}\beta^t\sum_{\substack{a\in A(j)\\ j\in S}}P_\pi(Y_t=j,\Delta_t=a|Y_0=i)r(j,a)$$

其中 $\beta\,(\in(0,1))$ 称为折扣因子.

下面将介绍折扣模型中最优策略与平稳策略的概念, 最优策略的基本理论及求解算法.

4.4.1 最优策略及其存在性

【定义 4.4】 *在折扣模型中, 设 $\pi^*\in\Pi$ 为一给定策略, 若对任给的 $\pi\in\Pi$, $i\in S$, 均有*

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

则称 π^* 关于折扣准则在 Π 中是 β 最优的, 简称为 β 最优策略.

可以证明, 在折扣模型中, 当状态集 S 及所有 $A(i),i\in S$ 均为有限时, 必存在一个 Markov 策略 π^*, 使得对任给的 $i\in S$ 及任意的 $\pi\in\Pi$, 均有

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

详细证明请参见文献 [27].

以上内容说明, 在全体策略类 Π 上寻求最优策略等价于在小得多的 Markov 策略类 Π_m^d 上寻求最优策略.

考虑到 $V_\beta(\pi,i)$ 的分量形式书写繁琐, 为简化后面的运算形式而引入矩阵符号.

设 $\pi=(f_0,f_1,\cdots,f_k,\cdots)\in\Pi_m^d$, 其中 f_k 表示 Markov 决策过程第 k 步时决策者所采取的行动规则, $k\in\{0,1,2,\cdots\}$, 该决策过程从 $t=k$ 到 $t=k+1$ 的一步转移概率显然是 f_k 的函数, 记为 $P(j|i,f_k(i))$, 其中 $i,j\in S=\{1,2,\cdots,l\}$. 并显然有 $P(j|i,f_k(i))\geqslant 0,\sum\limits_{j\in S}P(j|i,f_k(i))=1$. 记

$$P(f_k)=(P(j|i,f_k(i)))_{l\times l}$$

同理, $t=k$ 时的收益也是 f_k 的函数, 可写成 $r(i,f_k(i)),i\in S$, 并记

$$r(f_k)=(r(1,f_k(1)),r(2,f_k(2)),\cdots,r(l,f_k(l)))^T$$

再引入矩阵符号 $P^n(\pi)$ 并定义

$$P^0(\pi)=I,\quad P^n(\pi)=P(f_0)P(f_1)\cdots P(f_{n-1})$$

则 $P^n(\pi)$ 表示决策者采用 Markov 策略 $\pi=(f_0,f_1,\cdots,f_k,\cdots)$ 后的系统 n 步转移概率矩阵.

注意到当 $\pi=(f_0,f_1,\cdots,f_k,\cdots)\in\Pi_m^d$ 时, 对任何时刻 t 及 t 时刻系统所处的状态 j, 决策者按照策略 π 所采取的行动规则应惟一确定, 因此折扣目标可写为

$$V_\beta(\pi,i)=\sum_{t=0}^{\infty}\beta^t\sum_{j\in S}P_\pi(Y_t=j|Y_0=i)r(j,f_t(j))$$

其中 $P_\pi\{Y_t=j|Y_0=i\}$ 表示决策者采用 Markov 策略 $\pi=(f_0,f_1,\cdots,f_k,\cdots)$ 使系统在 $t=0$ 时刻从状态 i 出发, 于时刻 t 转移到 j 的 t 步转移概率. 于是, 用矩阵与向量的符号来书写, 折扣目标又可写为

$$V_\beta(\pi)=\sum_{t=0}^{\infty}\beta^tP^t(\pi)r(f_t)$$

4.4.2 平稳策略及其性质

【定义 4.5】 设 $\pi=(f_0,f_1,\cdots)\in\Pi_m^d$, 若对每个 $t\in\{0,1,2,\cdots\}$ 均有 $f_t\equiv f$, 则称为 (确定性) 平稳策略, 记做 f_0^∞, 全体平稳策略所成之集合记做 Π_s^d, 称它为平稳策略类.

显然有 $\Pi_s^d\subset\Pi_m^d\subset\Pi_m\subset\Pi$.

当采用平稳策略时, 折扣目标可写成

$$V_\beta(f^\infty)=\sum_{t=0}^{\infty}\beta^tP^t(\pi)r(f)$$

下面介绍平稳策略的一个重要性质.

【定理 4.2】 在折扣模型中, $\beta\in(0,1)$, $f\in F$, $P(f)=(P(j|i,f(i)))_{l\times l}$, 则有

(1) 矩阵 $I-\beta P(f)$ 的逆矩阵存在, 且有

$$\big[I-\beta P(f)\big]^{-1}=\sum_{k=0}^{\infty}\beta^kP^k(f)\tag{4.5}$$

(2) $V_\beta(f^\infty)=r(f)+\beta P(f)V_\beta(f^\infty)$, 其中 $V_\beta(f^\infty)$ 为采用平稳策略时的折扣目标函数, 且 $V_\beta(f^\infty)$ 存在且惟一.

对任给决策函数 $f\in F$, 仿照下式定义从 l 维列向量到 l 维列向量的变换 T_f

$$V_\beta(f^\infty)=r(f)+\beta P(f)V_\beta(f^\infty) \tag{4.6}$$

$$T_fV=r(f)+\beta P(f)V \tag{4.7}$$

对变换 T_f, 有如下性质:

【定理 4.3】 设 $f\in F$, T_f 为由 (4.7) 式定义的变换, V 与 V' 是两个 l 维列向量.

(1) 若 $V\geqslant V'$, 则有 $T_fV\geqslant T_fV'$, 即变换 T_f 是单调的.

(2) 若 $T_fV\geqslant V$, 则 $V_\beta(f^\infty)\geqslant V$, $V_\beta(f^\infty)$ 为采用平稳策略时的折扣目标. 同时, 若将 "$\geqslant$"改为 ">", 类似结论仍成立.

【定理 4.4】 设 π^* 为一 Markov 策略, 即 $\pi^*=(f_0,f_1,f_2,\cdots)\in\Pi_m^d$. 若 π^* 是 β 最优的, 则 π^* 在 $t=0$ 时刻的决策规则 (函数) f_0 所构成的平稳策略 f_0^∞ 对同一 β 也是最优的.

综合定理 4.1 与定理 4.4 的结论可得出如下事实: 在全体策略类 Π 上寻求最优策略, 等价于在平稳策略类 Π_s^d 上寻求最优策略. 因为在 Π_s^d 上所获得的 β 最优平稳策略, 在全体策略类 Π 上, 对同一 β 来说, 它同样是最优的. 考虑到当状态集 S 为有限以及所有 $A(i),i\in S$ 均为有限的假设下, Π_s^d 仅包含有限个不同的元素, 或仅有有限个平稳策略, 这就使得寻求最优策略的问题大为简化.

4.4.3 策略迭代法

【定理 4.5】 设一个给定的策略 $\pi^*\in\Pi$, 若对所有的 $f\in F$, 均有

$$T_fV_\beta(\pi^*)\leqslant V_\beta(\pi^*)$$

则 π^* 是 β 最优策略, 其中 $T_fV_\beta(\pi^*)=r(f)+\beta P(f)V_\beta(\pi^*)$.

利用定理 4.3 (2) 及定理 4.5 的结论可得如下策略迭代法的算法步骤:

第 1 步, 策略求值运算.

任取一个决策规则 $f\in F, i\in S=\{1,2,\cdots,l\}$, 求解如下 l 个线性方程组:

$$r(i,f(i))+\beta\sum_{j\in S}P(j|i,f(i))V(j)=V(i) \tag{4.8}$$

或

$$V=r(f)+\beta Q(f)V \tag{4.9}$$

其解 $V(i)=V_\beta(f^\infty,i)$.

第 2 步, 策略改进运算.

将第 1 步求得的 $V(i)(i\in S)$ 代入 (4.10) 式, 以寻求一个新的决策函数 $g=(g(1),g(2),\cdots,g(l))\in F$, 使其各分量分别满足下述关系式:

$$\begin{aligned}&\max_{a\in A(i)}\{r(i,a)+\beta\sum_{j\in S}P(j|i,a)V(j)\}\\&\quad=r(i,g(i))+\beta\sum_{j\in S}P(j|i,g(i))V(j)\\&\quad\geqslant r(i,f(i))+\beta\sum_{j\in S}P(j|i,f(i))V(j)\quad(i\in S)\end{aligned}\tag{4.10}$$

注意, 若同时有几个 a 使 (4.10) 式左端达最大, 则可任取其一作为 $g(i)(i\in S)$.

第 3 步, 终止规则.

若对所有的 $i\in S$, (4.10) 式均成立等式, 则终止计算, 并有结论: f^∞ 为 β 最优策略; 若至少存在一个 $i\in S$, 使 (4.10) 式成立严格不等式, 则以 g 代替 f, 并转入第 1 步, 此时有结论 $V_\beta(g^\infty)>V_\beta(f^\infty)$.

下面来说明上述算法步骤的原理. 对于任何一个决策规则 $f\in F$, 由算法第 2 步所定出的 g, 若按矩阵、向量符号书写可写为

$$r(g)+\beta P(g)V_\beta(f^\infty)\geqslant r(f)+\beta P(f)V_\beta(f^\infty)$$

由 (4.6) 式, 有

$$V_\beta(f^\infty)=r(f)+\beta P(f)V_\beta(f^\infty)$$

及 (4.7) 式, 有

$$T_gV_\beta(f^\infty)=r(g)+\beta P(g)V_\beta(f^\infty)$$

于是, 可得到

$$T_gV_\beta(f^\infty)\geqslant V_\beta(f^\infty)\tag{4.11}$$

由定理 4.3 (2) 知, 有

$$V_\beta(g^\infty)\geqslant V_\beta(f^\infty)$$

即经第 2 步所得的 g^∞ 至少是与 f^∞ 一样好的策略. 现分两种情况来讨论.

(1) 若 (4.11) 式等号成立, 则由 (4.10) 式, 对任给的 $h\in F$, 必有

$$r(h)+\beta P(h)V_\beta(f^\infty)\leqslant r(g)+\beta P(g)V_\beta(f^\infty)=V_\beta(f^\infty)$$

此即为

$$T_hV_\beta(f^\infty)\leqslant V_\beta(f^\infty)\quad(h\in F)$$

由定理 4.5 知 f^∞ 是 β 最优策略.

(2) 若 (4.11) 式严格不等号成立, 即有

$$T_\beta V_\beta(f^\infty) > V_\beta(f^\infty)$$

则由定理 4.3(2) 知有 $V_\beta(g^\infty) > V_\beta(f^\infty)$, 即 g^∞ 是比 f^∞ 更好的策略, 这种策略得到改进. 根据算法步骤, 将转入第 1 步, 并重复上述计算, 直到程序终止. 其中需说明的是, 由于 F 为有限集, 而每次迭代都实现严格改进, 因此不会出现循环现象, 即经过有限次迭代后, 将无法再做改进, 则根据前述论证, 此时的 f^∞ 必定在全体策略类 Π 上是 β 折扣最优的.

【例 4.4】 设有一工厂为市场生产某种产品. 每年年初对产品的销售情况进行一次检查, 其可能结果有两种: 销路好 (记为状态 1) 和销路差 (记为状态 2). 若销路好一年可获利 6 千元;若销路差一年要亏本 3 千元. 在每个状态工厂管理人员采用的行动均有两个: 不登广告 (记做 b) 或登广告 (记做 c). 若不登广告, 自然无广告费; 若登广告, 一年要花 2 千元广告费. 对于今年的各种状态及所采取的行动, 由于各种随机因素的干扰, 转为下年初的状态概率及相应状态需花费的费用见表 4.2. 工厂希望考虑长期折扣期望收益, 取折扣因子 $\beta = 0.9$, 求最优策略及其最优值函数 (计算取两位小数). 用策略迭代法求解此 MDP 问题.

表 4.2 状态转移概率及费用表

状态 i	行动	转移概率		报酬 (千元)
	$a = f(i)$	$P(1\mid i, f(i))$	$P(2\mid i, f(i))$	$r(a, i)$
1	b	0.5	0.5	6
	c	0.8	0.2	4
2	b	0.4	0.6	-3
	c	0.7	0.3	-5

解 由设知, 状态集 $S = \{1, 2\}$, 行动集 $A(1) = A(2) = \{b, c\}$. 该 Markov 决策过程的决策规则共有四个, 它们分别是

$$f = (f(1), f(2)) = (b, b), \quad g = (g(1), g(2)) = (b, c)$$
$$h = (h(1), h(2)) = (c, b), \quad \varphi = (\varphi(1), \varphi(2)) = (c, c)$$

(1) 任取一决策函数 $f = (f(1), f(2)) = (b, b)$ (实际上 $V_{0.9}(f^\infty)$ 是最差的折扣目标值) 做策略求值运算, 即解线性方程组

$$\begin{cases} 6 + 0.9[0.5V_1(1) + 0.5V_1(2)] = V_1(1) \\ -3 + 0.9[0.4V_1(1) + 0.6V_1(2)] = V_1(2) \end{cases}$$

解得

$$V_1 = (V_1(1), V_1(2)) = (V_{0.9}(f^\infty, 1), V_{0.9}(f^\infty, 2)) \approx (15.49, 5.60)$$

(2) 将上述计算得到的 V_1 代入 (4.10) 式, 以求解新的决策函数 $g_1 = (g_1(1), g_1(2))$. 注意到 $A(1) = \{b, c\}$, 故 (4.10) 式当 $i = 1$ 时, 取 $\beta = 0.9$, 有

$$\begin{aligned}&\max\{r(1,b) + \beta\sum_{j=1}^{2} P(j|1,b)V_1(j), r(1,c) + \beta\sum_{j=1}^{2} P(j|1,c)V_1(j)\}\\ &\quad = \max\{15.49, 16.16\} = 16.16\end{aligned}$$

故取 $g_1(1) = c$.

由于 $A(2) = \{b, c\}$, 故 (4.10) 式当 $i = 2$ 时, 取 $\beta = 0.9$, 有

$$\begin{aligned}&\max\{r(2,b) + \beta\sum_{j=1}^{2} P(j|2,b)V_1(j), r(2,c) + \beta\sum_{j=1}^{2} P(j|2,c)V_1(j)\}\\ &\quad = \max\{5.60, 7.52\} = 7.52\end{aligned}$$

故取 $g_1(2) = c$. 此时显然有 $V_{0.9}(g_1^\infty) > V_{0.9}(f^\infty)$.

(3) 以 $g_1 = (g_1(1), g_1(2)) = (c, c)$ 代替 f 转入第 1 步作策略求值运算, 即解下述线性方程组:

$$\begin{bmatrix} r(1,c) \\ r(2,c) \end{bmatrix} + \beta \begin{bmatrix} P(1|1,c) & P(2|1,c) \\ P(1|2,c) & P(2|2,c) \end{bmatrix} \begin{bmatrix} V_2(1) \\ V_2(2) \end{bmatrix} = \begin{bmatrix} V_2(1) \\ V_2(2) \end{bmatrix}$$

或有

$$\begin{bmatrix} 4 \\ 5 \end{bmatrix} + 0.9 \begin{bmatrix} 0.8 & 0.2 \\ 0.7 & 0.3 \end{bmatrix} \begin{bmatrix} V_2(1) \\ V_2(2) \end{bmatrix} = \begin{bmatrix} V_2(1) \\ V_2(2) \end{bmatrix}$$

求解得

$$V_2 = (V_2(1), V_2(2)) = (22.20, 12.31)$$

(4) 将上述 V_2 代入 (4.10) 式求解 $g_2 = (g_2(1), g_2(2))$. 与上同理, 由于 $A(1) = \{b, c\}$, 故 (4.10) 式当 $i = 1$ 时, 有

$$\max\{r(1,b) + \beta\sum_{j=1}^{2} P(j|1,b)V_2(j), r(1,c) + \beta\sum_{j=1}^{2} P(j|1,c)V_2(j)\}$$

$$= \max\{21.54, 22.20\} = 22.20$$

故取 $g_2(1) = c$. 类似地, 由于 $A(2) = \{b, c\}$, 故 (4.10) 式当 $i = 2$ 时有

$$\max\{r(2,b) + \beta\sum_{j=1}^{2} P(j|2,b)V_2(j), r(2,c) + \beta\sum_{j=1}^{2} P(j|2,c)V_2(j)\}$$
$$= \max\{11.64, 12.31\} = 12.31$$

故取 $g_2(2) = c$.

注意到 $g_2 = g_1 = (c, c)$, 这说明 g_1^∞ 已无法再作改进, 满足算法终止计算条件. 故 $g_2^\infty = g_1^\infty$ 为 $\beta = 0.9$ 之最优平稳策略.

相应的最优值函数为

$$V_{0.9}(g_2^\infty) = (V_{0.9}(g_1^\infty, 1), V_{0.9}(g_2^\infty, 2)) = (22.20, 12.31)$$

一般来说, 当状态空间 S 不很大时, 直接利用策略迭代算法来确定最优平稳策略, 但当状态空间 S 较大时, 需要解 l 个未知量的 l 个线性方程组, 计算量大, 是比较麻烦的.

4.4.4 逐次逼近法

最后, 只是简单地将逐次逼近法的步骤叙述如下, 有关的详细的介绍请参见文献 [5, 26].

第 1 步, 取一 l 维向量 $V_0 = 0$ 或 $V_0(i) = \max\limits_{a\in A(i)} r(i,a)$ $(i \in S)$.

第 2 步, 归纳地定义向量序列 $\{V_n\}$

$$V_{n+1} = TV_n = \max_{f\in F}\{r(f) + \beta P(f)V_n\} = r(f_n) + \beta P(f_n)V_n$$

此中的 max 是按分量分别取的, 即有

$$\begin{aligned} V_{n+1}(i) &= \max_{f(i)\in A(i)}\{r(i, f(i)) + \beta\sum_{j\in S} P(j|i, f(i))V_n(j)\} \\ &= r(i, f_n(i)) + \beta\sum_{j\in S} P(j|i, f_n(i))V_n(j) \quad (i\in S) \end{aligned} \tag{4.12}$$

由于 $A(i)$ 均为有限数, 故 f_n 一定存在.

初始向量 V_0 的选取将影响迭代所需要的步数. 因此在采用逐次逼近法解决实际问题时, 应根据已有的经验, 尽量选取较优的向量作为 V_0. 若无先验知识

可用, 通常可取 $V_0 = 0$ 或取 $V_0(i) = \max\limits_{a\in A(i)} r(i,a)(i\in S)$. 若取后者, 则按逐次逼近法经第 n 次迭代得到的 V_n 正好是从 0 到 n 周期内获得的最优折扣期望总报酬.

另外需要指出的是: 一般来说, 逐次逼近法并不提供一个得到最优策略的有限步迭代算法. 事实上, 最优值函数一般仅是逼近, 而不一定能达到, 但是经 n 次迭代后的值向量 V_n 与最优值函数 V_β^* 之间的误差, 是可以根据已有公式得到粗略估计的.

【例 4.5】 利用逐次逼近法求解例 4.4 中的 MDP 问题, 取 $\beta = 0.9$, 并计算 $n = 57$ 时的迭代值向量以及作相应的误差估计.

解 根据逐次逼近法的算法步骤, 取 $V_0 \equiv 0$, 按 (4.12) 式做一次迭代运算

$$\begin{aligned} V_1(1) &= \max_{a\in A(1)} \{r(1,a) + 0.9\sum_{j=1}^{2} P(j|1,a)V_0(j)\} \\ &= \max\{r(1,b), r(1,c)\} = r(1,b) = 6 \end{aligned}$$

达到右边最大的行动是 b, 故取 $f_1(1) = b$, 类似地, 有

$$\begin{aligned} V_1(2) &= \max_{a\in A(2)} \{r(2,a) + 0.9\sum_{j=1}^{2} P(j|2,a)V_0(j)\} \\ &= \max\{r(2,b), r(2,c)\} = r(2,b) = -3 \end{aligned}$$

达到右边最大的行动是 b, 故取 $f_1(2) = b$, 然后做第二次迭代运算

$$\begin{aligned} V_2(1) &= \max_{a\in A(1)} \{r(1,a) + 0.9\sum_{j=1}^{2} P(j|1,a)V_1(j)\} \\ &= \max\{r(1,b) + 0.9(0.5\times 6 + 0.5\times(-3)), \\ &\qquad r(1,c) + 0.9(0.8\times 6 + 0.2\times(-3))\} \\ &= \max\{6 + 0.9\times 0.5, 4 + 0.9\times 4.2\} \\ &= \max\{7.35, 7.78\} = 7.78 \end{aligned}$$

故取 $f_2(1) = c$, 类似地, 有

$$V_2(2) = \max_{a\in A(2)} \{r(2,a) + 0.9\sum_{j=1}^{2} P(j|2,a)V_1(j)\}$$

$$= \max\{r(2,b) + 0.9\sum_{j=1}^{2} P(j|2,b)V_1(j), r(2,c) + 0.9\sum_{j=1}^{2} P(j|2,c)V_1(j)\}$$
$$= \max\{-3 + 0.9(0.4 \times 6 + 0.6 \times (-3)), -5 + 0.9(0.7 \times 6 + 0.3 \times (-3))\}$$
$$= \max\{-2.946, -2.03\} = -2.03$$

故取 $f_2(2) = c$. 重复上述迭代计算, 将 $n = 57$ 次迭代计算结果列成表 4.3. 不难

表 4.3 逐次逼近法的迭代步骤

n	$V_n(1)$	$V_n(2)$	$f_{n-1}(1)$	$f_{n-1}(2)$	n	$V_n(1)$	$V_n(2)$	$f_{n-1}(1)$	$f_{n-1}(2)$
0	0	0			15	18.55	8.66	c	c
1	6	-3	b	b	20	20.05	10.16	c	c
2	7.78	-2.03	c	c	25	20.93	11.04	c	c
3	9.24	-0.65	c	c	30	21.44	11.55	c	c
4	10.54	0.65	c	c	35	21.76	11.87	c	c
5	11.71	1.82	c	c	40	21.94	12.05	c	c
6	12.76	2.87	c	c	45	22.06	12.17	c	c
7	13.70	3.81	c	c	50	22.11	12.22	c	c
8	14.55	4.66	c	c	55	22.16	12.26	c	c
9	15.32	5.42	c	c	57	22.16	12.27	c	c
10	16.01	6.12	c	c					

验证, V_{57} 还不满足最优方程 $TV = V$, 即 V_{57} 还不是最优值函数. 因此还需要继续进行迭代. 有关误差的进一步估计可参见文献 [26] 等.

可以看出, 策略迭代法当状态空间 S 所包含的元素个数 l 不太大时, 是一种有效的算法, 它比逐次逼近法所需的计算量要小得多. 但当 l 大时, 由于每次迭代策略迭代法要解 l 个未知量的 l 个线性方程组, 所需计算量甚多, 而逐次逼近法由于迭代规则简单, 很适合计算机计算. 因而有它的优越性. 因次采用策略迭代法与逐次迭代法的混合计算法将更有效. 借助于最优值函数的逐次逼近法可以在策略迭代法中选择一个比较理想的初始决策函数, 有关具体的做法请参见文献 [19].

最后, 如前所述, 若考虑使长期内每单位时间的平均期望报酬达到最大, 采用平均准则作为目标时, 此时的模型称为平均准则 Markov 决策规划, 在此不再作介绍, 有兴趣的读者可参见相关文献.

习 题

4.1 构造一个 Markov 决策规划的应用问题, 并指出其五元体中各元素在应用问题中的基本含义.

4.2 设某制造商生产并销售一种产品. 把月初销售状况分成好、中、差三个档次. 制造商可以根据月初销售情况采取不做广告或做广告两种措施. 为简单起见, 取状态空间 $S=\{1,2,3\}$, 其中状态 1,2,3 分别表示月初的销售状况为好、中、差. 对每一状态 i $(i=1,2,3)$, 均有行动集 $A(i)=\{1,2\}$, 其中 1 表示不做广告, 2 表示做广告. 由历史资料知, 对应于不同行动的转移概率矩阵分别为

$$P(1)=\begin{bmatrix}0.2 & 0.5 & 0.3\\ 0 & 0.2 & 0.8\\ 0 & 0 & 1\end{bmatrix},\quad P(2)=\begin{bmatrix}0.5 & 0.4 & 0.1\\ 0.1 & 0.6 & 0.3\\ 0.05 & 0.4 & 0.55\end{bmatrix}$$

每月利润依赖于月初销售状况及采取什么措施, 可用收益向量分别表示为

$$r(1)=(7,5,-1)^T,\quad r(2)=(5,4,2)^T$$

假设商品的营销周期仅为三个月. 问在每个月初应如何根据当时的销售情况确定该月是否要做广告, 以使这三个月内尽可能多获利.

4.3 考虑下述具有两个状态与行动的有限阶段的 Markov 决策问题. 总阶段数 $N+1=4$, 状态与行动均以 1 与 2 记之. 以行动 1 与 2 为参数的转移概率矩阵分别为

$$P(1)=\begin{bmatrix}\frac{1}{2} & \frac{1}{2}\\ \frac{2}{3} & \frac{1}{3}\end{bmatrix},\quad P(2)=\begin{bmatrix}\frac{1}{4} & \frac{3}{4}\\ \frac{1}{3} & \frac{2}{3}\end{bmatrix}$$

报酬函数 $r(i,a)$ 的四种可能取值为

$$r(1,1)=5,\quad r(1,2)=4,\quad r(2,1)=2,\quad r(2,2)=3$$

最终报酬为 $R(1)=2, R(2)=1$. 试求最优策略.

4.4 考虑一台运行的设备, 周期性地观察其状态, 设每次观察使该设备必处于如下四种状态之一: 0 设备完好; 1 设备稍微磨损, 仍可用; 2 设备严重磨损, 仍可用; 3 设备已损坏, 不能再用.