Modelos de Decisión en Ambientes Inciertos

(Apunte de Clases para el Curso Investigación Operativa IN44A)

Departamento de Ingeniería Industrial - Universidad de Chile

René A. Caldentey Susana V. Mondschein [1]

Enero, 1999

[1]La presente es una versión preliminar de este apunte docente, el cual se encuentra en construcción. Los autores agradecen los comentarios y correcciones de eventuales errores que aún permanezcan en el texto, los cuales pueden ser comunicados a smondsch@dii.uchile.cl, rcaldent@mit.edu o hawad@dii.uchile.cl

Capítulo 6

Modelos de Decisión Markovianos

6.1 Introducción

Hasta ahora nuestro rol en el estudio de procesos markovianos ha sido puramente descriptivo: describimos primero la evolución en probabilidad de un sistema modelable como una cadena de Markov, y después el valor esperado del beneficio acumulado, una vez que habíamos construido una estructura de beneficios sobre la cadena en cuestión. Sin embargo en las situaciones interesantes de la vida real no somos meros espectadores, sino que podemos tomar decisiones que afecten el comportamiento de los sistemas que estamos estudiando.

Consideremos por ejemplo una máquina productiva sujeta a posibles fallas, y a la cual se le pueden hacer distintos tipos de mantenciones preventivas a distintos costos. Supongamos que la ocurrencia de las fallas de la máquina responde a un proceso estocástico que se ve afectado por las mantenciones realizadas. En semejante situación desearemos escoger los tipos de mantención preventiva a realizar y los instantes en que conviene efectuarlas, de tal forma de maximizar los beneficios asociados a la operación del equipo.

El problema anterior es lo que llamaríamos un problema de decisión estocástico. En este capítulo estudiaremos un caso particular de problemas de decisión estocásticos y que son aquéllos en que el proceso estocástico corresponde a una cadena de Markov.

6.2 Definiciones

Consideremos un sistema cuya evolución en el tiempo está contenida al interior de un conjunto de estados finito $E = \{E_1, E_2, .., E_r\}$ y un cierto conjunto A que denominaremos conjunto de acciones posibles. Para todo $a \in A$ el par (E, a) representa una cadena de Markov finita y homogénea con matriz de probabilidades de transición de un período $P(a) = [p_{ij}(a)]$. Si además se conocen matrices de beneficio $R(a) = [r_{ij}(a)]\, a \in A$ en donde $r_{ij}(a)$ representa el beneficio de evolucionar del estado E_i al estado E_j durante un período usando la política a entonces el trío $(E, a, R(a))$ constituye a una cadena de Markov con beneficio.

6.3 Modelo de Horizonte Finito

El problema que nos interesa resolver en esta sección es determinar la secuencia de acciones que se deben tomar de tal forma de maximizar el valor esperado del beneficio acumulado durante n períodos. Dado que no conocemos de antemano cuál va a ser la evolución del sistema, no podemos decir a priori cuál es la acción que se debe elegir en cada período, sino que debemos entregar una regla de decisión que nos indique qué acciones tomar en función de la trayectoria (de estados) que siga el sistema. Ahora bien, como estamos limitados a reglas de decisión no anticipativas, la acción elegida cuando faltan n períodos para el final sólo puede depender del estado del sistema en ese momento y de la trayectoria previa (no puede depender de los estados que va a alcanzar el sistema después). Más aún, la condición de Markov nos indica que la información de la trayectoria previa es redundante, pues toda la información relevante para describir en probabilidad la evolución futura del sistema se encuentra en su estado actual. Así, lo que buscamos es una regla de decisión que nos indique qué acción tomar cuando faltan n períodos para el final del horizonte en función del estado del sistema en ese momento. Vale decir buscamos una función $s : E \times I\!N \longrightarrow A$, que para cada período n y cada posible estado E_i nos indique la mejor acción a seguir, $a = s(E_i, n)$.

Dada la estructura de nuestro problema resulta natural abordarlo con una formulación de programación dinámica. Definamos $V_k(i)$ como el beneficio esperado máximo si el sistema se encuentra actualmente en el estado E_i, faltan k períodos para el final del horizonte de planificación y en cada período se toman decisiones óptimas . La construcción de $V_k(i)$ (y de $s(E_i, k)$) se puede realizar recursivamente mediante la relación:

$$V_k(i) = \max_{a \in A} \left\{ \sum_{j=1}^{r} [r_{ij}(a) + V_{k-1}(j)] \cdot p_{ij}(a) \right\} \tag{6.1}$$

en donde $V_0(i)$ corresponde al valor residual de terminar la evolución en el estado E_i. La solución del problema anterior se realiza mediante las técnicas usuales de programación dinámica.

6.4 Modelo de Horizonte Infinito

La resolución del problema para el caso de horizonte infinito difiere considerablemente a la del caso finito debido a que las técnicas de programación dinámica ya no son aplicables. Además, en muchos casos el beneficio acumulado esperado será una función divergente en el largo del horizonte de planificación, de modo que puede no tener sentido preguntarse por las acciones que maximizan el beneficio acumulado esperado en el largo plazo. En tales casos resulta un mejor indicador de desempeño la tasa de crecimiento del beneficio acumulado esperado: buscaremos las acciones o reglas de decisión que maximicen el beneficio esperado por unidad de tiempo: $\lim_{k \longrightarrow \infty} \frac{V_k(i)}{k}$.

6.4.1 Políticas Estacionarias

Supongamos que en un período dado el sistema se encuentra en el estado E_i y somos capaces de determinar que la acción óptima es $a \in A$. Si k períodos más tarde el sistema vuelve al estado E_i, ¿qué acción convendrá elegir ahora? Dado que el horizonte de planificación es infinito, el problema a resolver es el mismo que k períodos antes, de modo que la acción óptima será nuevamente a. De esta forma vemos que, a diferencia del caso con horizonte finito, la acción óptima a seguir si el sistema se encuentra en E_i es independiente del período y depende exclusivamente del estado del sistema. Es por ello que la solución al modelo de horizonte infinito es lo que llamamos una *política estacionaria*.

Definición 6.1 *Una* Política Estacionaria *es una función* $s : E \to A$ *que a cada estado* $E_i \in E$ *asocia una acción* $s(E_i) \in A$.

Designemos por $\mathcal{S} = \{s : E \to A\}$ el conjunto de todas las políticas estacionarias posibles. Si el conjunto de acciones es finito, la cardinalidad de $\mathcal{S}$, $\|\mathcal{S}\| = \|A\|^{\|E\|}$ representa, como veremos más adelante, una medida de la dificultad para resolver un problema de horizonte infinito.

Ahora bien, toda política estacionaria s tiene asociada una matriz de probabilidades de transición en un período $P^s = \left[p^s_{ij}\right]$ tal que la fila i-ésima de P^s es igual a la fila i-ésima de $P(s(E_i))$. De la misma forma toda política estacionaria tiene asociada una matriz de beneficios $R^s = r^s_{ij}$ donde la fila i-ésima de R^s es exactamente igual a la fila i-ésima de $R(s(E_i))$.

De esta forma para toda política estacionaria s el trío (E, P^s, R^s) representa una cadena de Markov con beneficios y resolver el problema en horizonte infinito equivale a determinar aquella política $s \in \mathcal{S}$ que maximice el beneficio esperado por unidad de tiempo en el largo plazo (o el beneficio acumulado esperado en el largo plazo, según cuál de ellos tenga sentido).

6.4.2 Caso Ergódico

Supongamos que $\forall s \in \mathcal{S}$ la cadena de Markov representada por P^s es ergódica. Luego definiendo V^s_n el vector de beneficios esperados si faltan n períodos para el final y la política estacionaria escogida es s (note que no hay ningún proceso de optimización involucrado) y utilizando los resultados del capítulo anterior se tiene que:

$$V^s_n = ng^s e + W^s + [P^s]^n (V_0 - W^s), \tag{6.2}$$

donde g^s es el beneficio esperado por transición en estado estacionario y W^s es el vector asintótico de beneficios relativos cuando la política estacionaria usada es s.

Resolver el problema en horizonte infinito equivale a:

$$\max_{s \in S} \left\{ \lim_{n \to \infty} \frac{V^s_n}{n} \right\} \tag{6.3}$$

Ahora bien, de (6.2) vemos que el comportamiento de $\lim_{n\to\infty} V_n^s$ está en general regulado por el término $ng^s e$ (salvo cuando $g^s = 0$). Por lo tanto resolver (6.3) equivale a determinar:

$$\max_{s\in S}\{g^s\} \tag{6.4}$$

Una forma de resolver (6.4) es determinar para cada política estacionaria s el valor de g^s y luego escoger aquella con mayor valor. El principal problema de este procedimiento radica en el número de políticas estacionarias que un problema puede tener ($\|\mathcal{S}\| = \|A\|^{\|E\|}$ si A es finito), por ejemplo un problema con 10 estados y 3 acciones posibles tiene $3^{10} = 59049$ políticas estacionarias.

Una caracterización del óptimo de (6.4) que resulta útil para resolver el problema es la dada en el siguiente resultado.

Teorema 6.1 *Si para cada política estacionaria s la cadena de Markov representada por P^s es ergódica entonces dadas $s, s^* \in \mathcal{S}$ se tiene*

$$\begin{aligned} r^{s^*} + P^{s^*}W^{s^*} \geq r^s + P^sW^{s^*} \quad &\Rightarrow \quad g^{s^*} \geq g^s, \\ r^{s^*} + P^{s^*}W^{s^*} \leq r^s + P^sW^{s^*} \quad &\Rightarrow \quad g^{s^*} \leq g^s, \end{aligned} \tag{6.5}$$

donde r^s, es el vector de beneficios esperados en una transición asociado a la política s, $r^s(i) = \sum_{j=1}^{r} r_{ij}^s \cdot p_{ij}^s$. W^s y g^s corresponden al beneficio esperado por transición en estado estacionario y al vector asintótico de beneficios relativos cuando la política estacionaria usada es s.

Corolario 6.1 *$s^* \in \mathcal{S}$ es óptima para (6.4) si y sólo si satisface:*

$$r^{s^*} + P^{s^*}W^{s^*} \geq r^s + P^sW^{s^*} \quad \forall s \in \mathcal{S} \tag{6.6}$$

La demostración se puede encontrar en Howard (1960), Gallager (1995) o Ross (1987), entre otros.

Es necesario destacar que en las desigualdades vectoriales de las ecuaciones (6.5) y (6.6) el cumplimiento de la condición para la componente i depende sólo de la acción tomada en el estado E_i por las políticas comparadas. El Teorema 6.1 sugiere la utilización de un algoritmo para encontrar la política estacionaria óptima, el que se conoce como algoritmo de Howard.

Algoritmo de Howard

1. Seleccionar una política estacionaria $\bar{s} \in \mathcal{S}$.

2. Calcular $W^{\bar{s}}$.

3. Construir la política estacionaria s de la siguiente forma:

$$s(E_i) = \arg\max_{a \in A} \left\{ r_i(a) + \sum_j p_{ij}(a) W_j^{\bar{s}} \right\}$$

4. Si s es igual a $\bar{s}$, entonces $\bar{s}$ es óptima (salir). Si no, reemplazar $\bar{s}$ por s y volver a 2.

Una pregunta natural que aparece al existir dos procedimientos de resolución de (6.4) es cuál de ellos requiere de menos esfuerzo; en este caso es fácil responder pues el primer procedimiento calcula todos los g^s y por tanto requiere en general más esfuerzo que el algoritmo de Howard. Ahora bien, qué tanto más rápido es el algoritmo de Howard respecto del método de enumeración no es tan directo de responder. En principio calcular para una política g^s tiene una complejidad comparable a calcular W^s. Luego si decimos que el método de enumeración tiene una complejidad de orden $\|S\|$, el algoritmo de Howard tiene una complejidad dada por el número de iteraciones que realiza.

6.5 Ejercicios

1. **Un modelo simple de inventarios.** Suponga que ud. es el encargado de planificación de la producción en la empresa Y, la cual fabrica y vende el producto X.

   La demanda por el producto X en una semana dada puede tomar sólo dos valores: 0 o 1 [unidades]. El comportamiento de la demanda es susceptible de ser modelado como una cadena de Markov: la probabilidad que la demanda sea igual a 1 en una semana dada es α si la demanda fue 1 la semana anterior y β si la demanda fue 0 la semana anterior, independiente de lo que haya ocurrido 2 o más semanas antes.

   El precio venta de una unidad de X es de A[\$] y el costo variable de producción es de C[\$] ($A > C$). Cada semana se puede fabricar a lo más 1 unidad de X. Además, cada vez que se inicia un ciclo productivo (entendido como una secuencia de semanas en todas las cuales se produce) se incurre en un costo de setup S[\$] ($S > 0$). A modo de ejemplo: si se produce X durante 4 semanas seguidas el costo de producción total es $4C + S$, mientras que si se produce durante 2 semanas, luego se detiene la producción una semana, y en seguida se produce durante 2 semanas más, el costo total de producción es $4C + 2S$.

   (a) ¿Qué información es relevante para tomar la decisión de producir o no en una semana cualquiera? Formule (no resuelva) un modelo de decisión markoviano que permita tomar esa decisión.

   (b) ¿Qué forma toma una política estacionaria para este problema? Dé un ejemplo de una política estacionaria.

(c) ¿Qué indicador utilizaría para decir que una política estacionaria es mejor que otra? Muestre que la política estacionaria "producir siempre" no es la óptima, argumentando que la siguiente política estacionaria es mejor que ella para algún valor de T: comenzar un ciclo productivo cada vez que el inventario caiga a 1 y continuar produciendo hasta que el inventario llegue a T. Si gusta apoye su argumento en el caso particular (y determinístico) en que $\alpha = 0$ y $\beta = 1$.

Bibliografía

[1] Howard, R. A. *Dynamic Programming and Markov Processes.* Wiley, New York, 1960.

[2] Gallager, Robert G. *Discrete Stochastic Processes.* Kluwer, Boston, 1995.

[3] Ross, Sheldon *Applied Probabilty Models with Optimization Applications.* Dover Books, 1992.