Boletín de Matemáticas
Nueva Serie, Volumen VIII No. 1 (2001), pp. 31–45

PROCESOS MARKOVIANOS DE DECISIÓN EN EL MODELAMIENTO DE AGENTES RACIONALES

YOFRE H. GARCÍA G. (*)

Resumen. Se presenta una síntesis teórica de los Procesos Markovianos de Decisión con horizonte infinito, junto con el método de iteración de valores. Se mencionan conceptos de Inteligencia Artificial (AI), como el concepto de agente, útiles en la formulación del problema de la rejilla.

Abstract. A theoretical synthesis of the Markov Decision Processes with infinite horizon, the value iteration method and some concepts of Artificial Intelligence (AI), that describe the simple reflex agent, useful in the modelling of the grid world problem, are presnted.

Palabras clave: Procesos Markovianos de decisión. MDP´s totalmente observables. Teoría de la decisión. Problemas de planificación totalmente observables. Inteligencia Artificial. Agentes racionales.

Key Words: Fully Observable Markov Decision Processes FOMDP´s. Decision Theory. Planning problems fully observables. Artificial Intelligence. Agents in AI.

1. Introducción

La toma secuencial de decisiones bajo condiciones de incertidumbre en el campo de la Inteligencia Artificial (AI), es un problema que ocupa parte importante de las investigaciones recientes que se desarrollan a partir de una visión más unificada de esta área, obtenida del enfoque basado en agentes [12]. El estudio y construcción de agentes capaces de actuar racionalmente con información incierta, se basa en herramientas de otras áreas como la Teoría de las Decisiones.

(*) Yofre H. García G., Departamento de Matemáticas, Universidad Nacional de Colombia. e-mail: ygarcia @matematicas.unal.edu.co
El autor agradece a la profesora Myriam Muñoz de Ozak por sus valiosas sugerencias a este documento.

Aquí se presentan los procesos Markovianos de decisión (MDP´s), como una alternativa para resolver problemas de decisión totalmente observables en IA, como el problema de la rejilla [12], [9], [14] y [10].

1.1 Notación. Si (X, τ) es un espacio métrico, $\mathfrak{B}(X)$ representa la menor $\sigma-$álgebra sobre X que contiene todos los abiertos que conforman a τ [3], decimos que X, junto con la colección $\mathfrak{B}(X)$, es un espacio de Borel. Un elemento $B \in \mathfrak{B}(X)$ se llama conjunto de Borel.

$B(X)$ es el espacio de Banach de todas las funciones medibles y acotadas a valor real sobre X (el espacio de estados), bajo la métrica del supremo.

Para una política $\delta \in \Delta$, P_x^δ representa la medida de probabilidad sobre X concentrada en $\{x\}$ [3], cuando se adopta la política δ. El valor esperado con respecto a la medida P_x^δ, lo denotamos por E_x^δ.

La expresión $A(x_i)$ representa el conjunto de acciones admisibles sobre un fenómeno aleatorio ó sistema estocástico [6], que se encuentra en el estado x_i.

2. Procesos Markovianos de Decisión (MDP´s)

Los MDP´s propuestos por Bellman [1] en 1957, con base en las ideas de Shapley [13], se han utilizado comúnmente en teoría de la decisión para el modelamiento de problemas de decisión secuencial sobre fenómenos aleatorios. Decisión secuencial en un fenómeno aleatorio es la elección de una acción en cada instante del tiempo.

En la formulación de un problema de decisión secuencial, se supone que un sistema o fenómeno cambia aleatoriamente de estado cuando se ejecutan sobre él acciones a través del tiempo, mientras el sistema proporciona recompensas. El problema de decisión secuencial consiste en determinar la sucesión de acciones que acumule la mayor cantidad de recompensas del sistema.

Una interpretación económica de las recompensas (como ganancias o costos), corresponde a la formulación clásica de estos problemas, donde la estructura probabilística se modela por medio de las cadenas de Markov estacionarias con parámetro de tiempo discreto, como se expone en [2]. Sin embargo, una mejor descripción de las probabilidades de transición en estos problemas se logra bajo la definición de núcleo estocástico:

Definición 1. *Sean X y Y espacios de Borel. Un núcleo estocástico sobre X dado Y, es una función $q(x \mid y)$ tal que, para cada $y \in Y$, $q(\cdot \mid y)$ es una medida de probabilidad sobre X, y para cada conjunto de Borel $B \in \mathfrak{B}(X)$, $q(B \mid \cdot)$ es una función medible de Y hacia* $[0, 1]$.

Con esta definición, la interacción en un problema de decisión secuencial está dada de la siguiente forma: Si el sistema está en el estado x_i, por medio de la función de decisión $f_{n-1} : X \to A(x_i)$ del instante $n-1$, se elige la mejor acción $a_i = f_{n-1}(x_i)$ del conjunto de acciones admisibles $A(x_i)$. Esta decisión determina las probabilidades de transición [2], del sistema $\Pi_{ij} = q(x_j \mid$

$x_i, f_{n-1}(x_i))$, descritas bajo el núcleo estocástico de X dado $X \times A(x_i)$. Π_{ij} es la probabilidad de que el sistema pase al estado x_j dado que se encuentra en el estado x_i y se ha elegido por medio de la función de decisión f_{n-1} la acción $f_{n-1}(x_i) \in A(x_i)$. Además, por medio de la función $r : X \times A \to \mathbb{R}$, el sistema proporciona la recompensa $r(x_i, f_{n-1}(x_i))$, es decir, la recompensa obtenida cuando el estado del sistema es x_i y se adopta la acción $f_{n-1}(x_i) \in A(x_i)$. Gráficamente, esto corresponde al esquema de la figura 1.1.

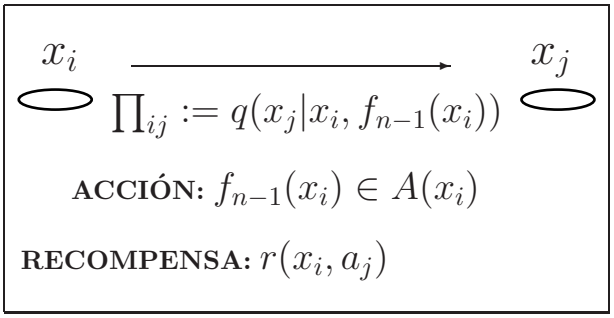

Figura 1.1

Para el modelamiento de estos problemas se usan los procesos Markovianos de decisión (MDP´s) estacionarios, definidos como sigue:

Definición 2. *Un Proceso Markoviano de decisión estacionario (MDP)* [5] *es una colección de objetos* (X, A, q, r) *donde,*

- X *es un espacio de Borel que denota el conjunto de posibles estados del sistema.*
- A *es también un espacio de Borel, pero que denota el conjunto de acciones del sistema.*
- q *es un núcleo estocástico sobre* X *dado* $X \times A$.
- r *es una función real medible sobre* $X \times A$, *conocida como la función de recompensas.*

Este modelo genera la sucesión de estados y acciones $(x_t, a_t)_{t \in T}$ del problema de decisión secuencial. Un MDP estacionario es de horizonte infinito si T es un conjunto infinito numerable. Comúnmente el conjunto T son los números naturales. La sucesión $(x_n)_{n \in \mathbb{N}}$ que se deriva de un MDP es una cadena de Markov estacionaria de parámetro de tiempo discreto. A la sucesión de funciones de decisión $(f_n)_{n \in \mathbb{N}}$ que definen a la sucesión de acciones $(a_n)_{n \in \mathbb{N}}$ se llama una política. Cuando la función de decisión que se aplica en cada instante del tiempo es la misma, decimos que $(f_n)_{n \in \mathbb{N}}$ es una política estacionaria, denotada por δ o simplemente por f. Al conjunto de políticas estacionarias lo denotamos

por Δ. Así, la elección de acciones en un proceso de decisión secuencial corresponde a escoger la política estacionaria más adecuada para que el sistema proporcione recompensas bajo algún criterio de optimización.

Un problema de decisión secuencial puede convertirse en un problema de optimización, conocido como problema de decisión de Markov, al definir un MDP estacionario y un criterio de optimización.

Un criterio de optimización resulta a partir de una función medible a valor real $v : \Delta \times X \to \mathbb{R}$, que depende del estado inicial del sistema y de la política estacionaria. Los criterios de optimización clásicos en teoría de la decisión para un MDP son: el criterio de recompensa total esperada con factor de descuento β, definido como

$$v(\delta, x) = E_x^{\delta}\left[\sum_{t=o}^{\infty} \beta^t r(x_t, a_t)\right], \ \beta \in (0, 1),$$

y el criterio de recompensa esperada promedio

$$J(\delta, x) = \liminf_{n\to\infty} E_x^{\delta} \frac{\left[\sum_{t=o}^{n-1} r(x_t, a_t)\right]}{n}, \ \beta \in (0, 1),$$

siempre y cuando r sea una función acotada.

En el criterio de recompensa total esperada con descuento, el valor $\beta \in (0, 1)$ representa el descuento secuencial de la recompensa obtenida del sistema, es decir, se especifica β de tal forma que el valor de una unidad de la recompensa en el tiempo $n = t$ se convierta para el tiempo $n = t + k$ en β^k. Con estos elementos en juego, si V define un criterio de optimización, el problema de elegir la mejor acción sobre el sistema en cada instante del tiempo, se traduce en encontrar la política estacionaria $\delta^* \in \Delta$ tal que

$$v(\delta^*, x) = \sup_{\delta\in\Delta} V(\delta, x), \text{ para todo } x \in X.$$

La política estacionaria δ^* que cumpla esta condición se llama política óptima. A la función $v^*(x) = \sup_{\delta\in\Delta} V(\delta, x)$, para todo $x \in X$ se le llama función óptima de recompensa.

Definición 3. *Una función $D : X \to K(A)$, donde $K(A)$ denota los subconjuntos compactos de A, es semicontinua por arriba si para cada subconjunto abierto A' de A, el conjunto $\{x : D(x) \subseteq A'\}$ es un abierto en X.*

Las condiciones básicas que garantizan la existencia y unicidad de la política óptima, y de la función óptima de recompensa, independientemente del criterio de optimización utilizado son las siguientes:

Condiciones básicas de optimización:

1. Para cada estado x, el conjunto de acciones admisibles $A(x)$, es un subconjunto compacto no vacío de A.

2. Para alguna constante R, $|r(x,a)| \leq R$ para todo $(x,a) \in X \times A(x)$, y además para cada $x \in X$, $r(x,\cdot)$ es una función semicontinua por arriba.
3. $\int v(y)q(dy \mid x,\cdot)$ es una función semicontinua por arriba para cada $x \in X$ y para cada función $v \in B(X)$.

La existencia de la política estacionaria se garantiza por el siguiente teorema [7]:

Teorema 1. *Sea $D : X \to K(A)$ una función Borel medible, y $v(x,a)$ una función real medible sobre el grafo de D, que es semicontinua por arriba en la variable $a \in D(x)$, para cada $x \in X$. Entonces:*

1. *Existe una función de decisión $f : X \to A$ a través de la función D, tal que*

$$v(x, f(x)) = \max_{a \in D(x)} V(x,a) \text{ para todo } x \in X,$$

y además, la función $v^*(x) = \max_{a \in D(x)} v(x,a)$, es Borel medible.

2. *Si D y v son semicontinuas por arriba y acotadas, entonces v^* es semicontinua por arriba y acotada.*■

La existencia y unicidad de la política óptima y la función de recompensa cuando se usa el criterio de recompensa total esperada con factor de descuento, se garantiza por el siguiente teorema [6]:

Teorema 2. *Bajo las condiciones básicas de optimización se tiene que:*
1. La función óptima de recompensa v^* es la única solución en $B(X)$, de la ecuación

$$v^*(x) = \max_{a \in A(x)} \left\{ r(x,a) + \beta \int_X v^*(y) q(dy \mid x,a) \right\}, \text{ para todo } x \in X. \tag{1}$$

2. Una política estacionaria , $\delta^* \in \Delta$, es óptima sí y sólo si $\delta^*(x)$ maximiza la parte derecha de la ecuación anterior, es decir

$$v^*(x) = \left\{ r(x, \delta^*(x)) + \beta \int_X v^*(y) q(dy \mid x, \delta^*(x)) \right\}, \text{ para todo } x \in X.$$

Para demostrar este teorema se construye el operador $T : B(X) \to B(X)$,

$$T(v(x)) = \max_{a \in A(x)} \left\{ r(x,a) + \beta \int_X v(y) q(dy \mid x,a) \right\}.$$

Este es un operador de contracción con módulo β, es decir

$$\|Tu - Tv\| \leq \beta \|u - v\|,$$

para todo $u, v \in B(X)$, cuando $\|\cdot\|$ es la norma del supremo entre funciones. Por el teorema de punto fijo de Banach [8], se garantiza que este operador tiene un punto fijo, es decir, $Tv^* = v^*$, para alguna $v^* \in B(X)$. De otro lado, la existencia y unicidad de la política óptima y de la función de recompensa,

cuando se adopta el criterio de recompensa esperada promedio es el siguiente [6]:

Teorema 3. *Supóngase que se tienen las condiciones de optimización, y que existe además una constante j^* y una función $v^* \in B(X)$, tal que:*

$$j^* + v^* = \max_{a \in A(x)} \left\{ r(x,a) + \beta \int_X v^*(y) q(dy \mid x,a) \right\}, \text{ para todo } x \in X. \tag{2}$$

Entonces,
1. $\sup_{\delta} J(\delta, x) \leq j^*$, *para todo* $x \in X$.
2. Si f^* es la política estacionaria tal que $f^*(x) \in A(x)$ maximiza la parte derecha de la ecuación

$$j^* + v^* = r(x, f^*(x)) + \beta \int_X v^*(y) q(dy \mid x, f^*(x)), \text{ para todo } x \in X,$$

entonces f^* es la política óptima y $J(x, f^*(x)) = j^*$ para todo $x \in X$.
3. Para cualquier política δ y cualquier $x \in X$,

$$\lim_{n \to \infty} n^{-1} \sum_{i=0}^{n-1} r(x_i, a_i) = j^*, \; P_x^{\delta} - \textit{casi siempre},$$

si y sólo si

$$\lim_{n \to \infty} n^{-1} \sum_{i=0}^{n-1} \phi(x_i, a_i) = j^*, \; P_x^{\delta} - \textit{casi siempre}, \tag{3}$$

donde $\phi(x,a)$ es una función sobre $X \times A$ definida por

$$\phi(x,a) := r(x,a) + \beta \int_X v^*(y) q(dy \mid x,a) - j^* - v^*(x).$$

4. Si δ satisface (3), para todo $x \in X$, entonces δ es óptima.

Con base en estos resultados, se han construido métodos para calcular la política óptima cuando los estados y las acciones forman conjuntos finitos. En este caso, las ecuaciones (1) y (2) toman la siguiente forma:

$$v^*(x) = \max_{a \in A(x)} \left\{ r(x,a) + \beta \sum_{y \in X} v^*(y) q(dy \mid x,a) \right\}, \text{ para todo } x \in X, \tag{4}$$

y

$$j^* + v^* = \max_{a \in A(x)} \left\{ r(x,a) + \beta \sum_{y \in X} v^*(y) q(dy \mid x,a) \right\}, \text{ para todo } x \in X. \tag{5}$$

El método de iteración de valores o de aproximación sucesiva cuando se considera el criterio de recompensa total esperada con descuento, consiste en aplicar la función recursiva:

$$v_{n+1}(x) = \max_{a \in A(x)} \left\{ r(x,a) + \beta \sum_{y \in X} v_n(y) q(dy \mid x, a) \right\}, \text{ para todo } x \in X,$$

hasta

$$\|v_n(x) - v_{n+1}(x)\| \leq \frac{\varepsilon(1-\beta)}{\beta}, \text{ para todo } x \in X.$$

La fracción $\frac{\varepsilon(1-\beta)}{\beta}$, se conoce como el error de Bellman. Cuando la diferencia entre las aproximaciones de la función de recompensa satisface esta cota, dada en función del descuento aplicado y del error admisible, puede probarse que se obtiene una política aproximadamente óptima [5].

Para usar este mismo método cuando se aplica el criterio de recompensa promedio, es necesario garantizar que las probabilidades de transición entre los estados se estabilizarán a lo largo del tiempo, independientemente de la acción que se escoja. Para describir esto, recordemos que en un MDP, al fijar una regla de decisión f, el proceso se convierte en una cadena de Markov estacionaria con parámetro de tiempo discreto. Esto nos permite aplicar algunos resultados sobre cadenas de Markov que describen el comportamiento asintótico de las probabilidades de transición. Denotaremos por Q_f a la matriz de transición entre los estados del proceso, cuando se aplica la regla de decisión f. Si Q_f representa una cadena de Markov estacionaria irreducible y no periódica [2], entonces, puede probarse que $\{Q_f\}^t \to Q_f^*$, donde Q_f^* es la matriz estocástica de las probabilidades de transición cuando han transcurrido un número infinito de periodos de tiempo. Con esta matriz y teniendo en cuenta que X es finito, podemos denotar la recompensa esperada promedio para todos los estados del MDP por medio del vector $g := Q_{f^*}^* \cdot r_{f^*}$, donde r_{f^*} es la función de recompensa cuando se fija la política estacionaria f^*.

Definición 4. *A la expresión $H_{Q_f} = (I - (Q_f - Q_f^*))^{-1}(I - Q_f)$, se le conoce como Bias de la matriz Q_f.*

A partir de los anteriores elementos, la forma vectorial de la función recursiva que se aplica en este caso es $v_n = H_{Q_f} \cdot r_f + n(Q_f^* r_f) + Q_f^* v_0$. Cuando en un MDP, las matrices Q_f provienen de una cadena de Markov estacionaria irreducible y no periódica para todas las políticas estacionarias f, se garantiza que a partir de la anterior expresión, se encuentra una aproximación a la recompensa total esperada promedio como un valor constante de la diferencia entre las iteraciones [11], es decir, $j^* = v_n - v_{n-1}$, siempre y cuando

$$\left| \max_{x \in X} \{v_n(x) - v_{n-1}(x)\} - \min_{x \in X} \{v_n(x) - v_{n-1}(x)\} \right| \leq \varepsilon.$$

3. Una aplicación de los MDP´s en Inteligencia Artificial (AI)

A partir del enfoque de la Inteligencia Artificial basado en agentes, se ha logrado dar una visión más unificada de esta área y formular con mayor generalidad muchos problemas clásicos que allí se han propuesto.

El objetivo principal de la AI bajo este enfoque es el diseño de agentes inteligentes.

Definición 5. *Un agente es todo aquello que puede considerarse que percibe su ambiente mediante sensores y que responde o actúa en tal ambiente mediante efectores* [12].

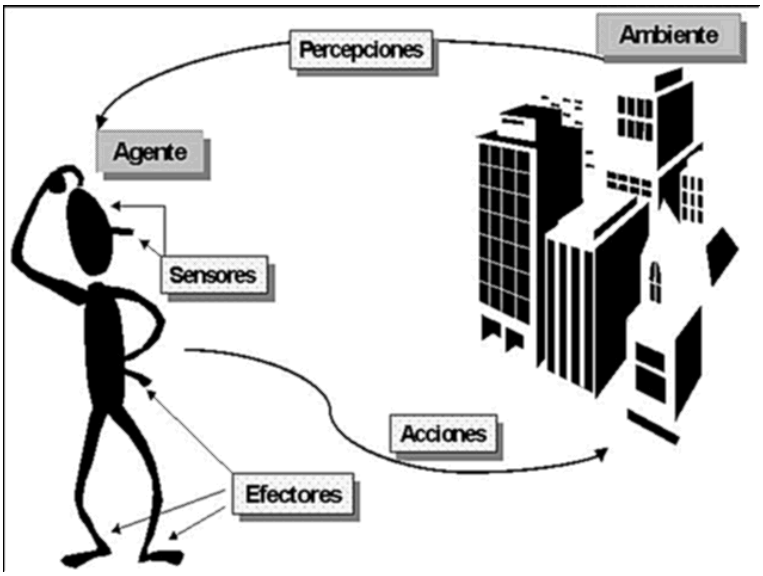

Figura 3.1. Esquema general de agente.

Un ejemplo de la interacción agente-ambiente se obtiene entre un virus informático y un sistema operativo. Los sensores y efectores de un virus informático son algoritmos que le permiten al virus identificar el tipo de sistema operativo, para luego ejecutar el algoritmo adecuado que pueda causar alguna alteración. Las percepciones y acciones de este agente son bits codificados. En el diseño de agentes inteligentes, es necesario dar una definición satisfactoria de lo que entenderá por una acción inteligente. Este concepto puede darse en función de las capacidades del ser humano o con base en un concepto ideal de inteligencia denominado racionalidad.

Definición 6. *Un agente es racional si hace lo correcto, donde lo correcto es aquello que contribuye a la ejecución de acciones que le permiten al agente obtener el máximo desempeño esperado en todos los casos posibles de secuencias de percepciones, basándose siempre en la información que recopila del ambiente y en todo su conocimiento incorporado.*

A diferencia del concepto de inteligencia en función de las capacidades humanas, por medio de la definición de agente racional, se facilita la intervención de otras áreas del conocimiento como la matemática y la ingeniería, en la construcción y diseño de agentes en (AI). Esta es una gran ganancia dentro de las metodologías de investigación en el campo de la AI, ya que en la actualidad, es más común que se propongan métodos y teorías sobre elementos de otras áreas con mayores bases sólidas y rigurosas, que lanzar nuevas hipótesis fruto de la intuición.

Además, este enfoque es mucho más general que el enfoque clásico de la AI basado en las "leyes del pensamiento" [12], que se centra en construir inferencias lógicamente correctas, utilizando elementos de la lógica de primer orden como herramienta esencial para representar el conocimiento y razonar con base en él. Sin embargo, en ambientes donde la información que percibe el agente no tiene el ciento por ciento de certeza, no hay garantía de que las acciones escogidas bajo razonamiento puramente lógico contribuyan al alcance de las metas y al buen desempeño del agente.

Para tratar de garantizar que un agente actúe racionalmente en ambientes con condiciones de incertidumbre, es necesario que el conocimiento incierto se represente con herramientas de la teoría de la probabilidad, y para que la selección de acciones influya directamente en la medida del desempeño del agente, se requiere adicionalmente de la lógica, junto con algunos elementos de la teoría de la utilidad.

La utilización de algunos tópicos de las áreas antes mencionadas, ha permitido el desarrollo de métodos de planificación para resolver problemas en AI bajo condiciones de incertidumbre y hacen parte de lo que actualmente se conoce como "decision theoretic planning" (DTP) [4].

Una de las principales herramientas de la DTP, que combina elementos de las teorías de la probabilidad y utilidad, son los procesos Markovianos de decisión. La aplicación de los MDP´s bajo esta nueva interpretación se ha extendido notablemente en (AI), a tal punto que las investigaciones actuales se dedican a explorar modelos avanzados de MDP´s, como los procesos Markovianos de decisión parcialmente observables (POMDP), y no observables (NOMDP) [4]. Los procesos Markovianos de decisión que se presentan aquí, se utilizan en este contexto para describir problemas de planificación bajo condiciones de incertidumbre totalmente observables, esto es cuando el agente conoce todos los estados del ambiente, una extensión de los procesos de decisión secuencial.

Los MDP´s utilizados para modelar este tipo de problemas, se llaman dentro de la DTP procesos Markovianos de decisión totalmente observables (FOMDP).

4. FOMDP en la construcción del agente de reflejo simple

En un problema de planificación bajo condiciones de incertidumbre en AI, la racionalidad de un agente dependerá de que en cada instante del tiempo, tome la mejor acción en el ambiente, de tal forma que contribuya a la acumulación del máximo desempeño esperado. Esta condición es la versión del principio de máxima utilidad esperada que se utiliza en teoría de la decisión y es el fundamento más básico para la construcción de los criterios de optimización [1]. El lugar que ocupan cada uno de los elementos que conforman un MDP en estos problemas es el siguiente:

El conjunto de estados del proceso está formado por todos los estados del ambiente. El conjunto de acciones se compone de todas las posibles acciones que puede emprender el agente a través de sus efectores. Las probabilidades condicionales que proporciona el núcleo estocástico se interpretan en este contexto como las probabilidades de transición entre los estados del ambiente cuando el agente racional elige una acción. La función de recompensas se interpreta casi siempre como los costos que produce el agente, de tal forma que el máximo desempeño esperado se obtiene de la minimización de la función total de recompensa . Esta medida del desempeño es desde luego el mecanismo para determinar si las acciones que elige el agente son en verdad racionales. Por lo tanto, la definición de esta función y su cálculo los efectúa un observador externo al problema de planificación. La característica especial de un problema de planificación bajo condiciones de incertidumbre totalmente observable, es que las percepciones del agente permiten identificar completamente el estado actual del ambiente para ejecutar sin problema la política óptima. Así, el agente racional construido a partir de un MDP debe estar dotado de sensores que le permitan reconocer el estado del ambiente, y efectores suficientes para la ejecución de la política estacionaria que forma su estructura interna y su único conocimiento del ambiente. Un agente con estas cualidades es el agente de reflejo simple, cuya estructura general es la que se muestra en la figura 4.1. Las reglas de condición-acción del Agente de Reflejo Simple corresponden a la política óptima. Las demás características que dan paso a la formulación de un FOMDP para resolver estos problemas, las posee el ambiente. Un ambiente con limitadas percepciones y acciones claramente distinguibles, donde la experiencia y actuación del agente se divide en episodios, corresponde a un ambiente discreto y episódico. Si el agente puede con sus sensores identificar el estado total del ambiente y no existe la posibilidad de que se modifique mientras el agente se encuentra deliberando en la elección de la acción, el ambiente es asequible y dinámico con relación al agente. Además, si las acciones emprendidas por el agente no determinan completamente el siguiente estado del ambiente, se

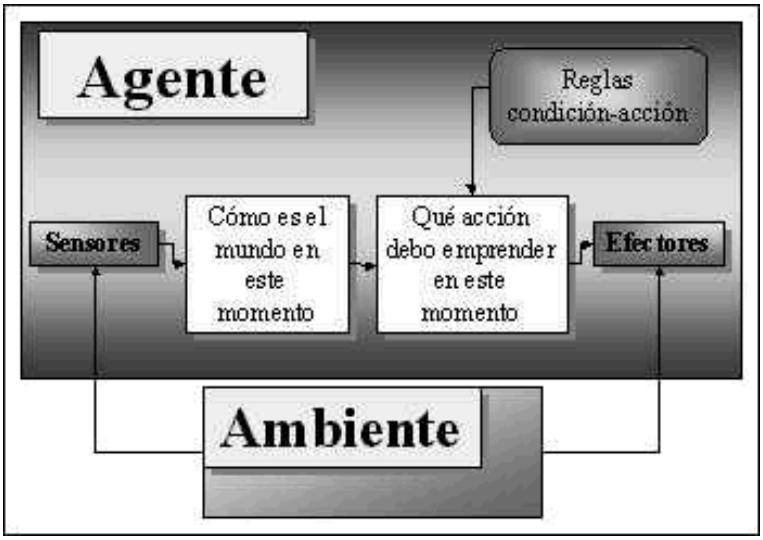

Figura 4.1. Estructura general del agente de reflejo simple.

dice que el ambiente es estocástico. Todas las anteriores cualidades de los ambientes deben estar presentes en la formulación de un FOMDP para construir las reglas de condición-acción que al ser ejecutadas por el agente de reflejo simple, garantizarán su máximo desempeño esperado en la solución de problemas de planificación totalmente observables bajo condiciones de incertidumbre.

5. El problema de la rejilla

Este problema clásico de incertidumbre en AI [9], [10] y [12], aparece de la siguiente situación: supongamos que dentro de un cuarto se observa el desplazamiento de un agente dotado de un sensor que le permite identificar claramente la posición que ocupa dentro del cuarto. Para facilitar los cálculos, supongamos que este cuarto está dividido en 12 posiciones tal como se representa en la figura 5.1, una de ellas ocupada por un objeto. Una vez que el agente inicia su desplazamiento genera una recompensa de -0.04 (costo de desplazamiento) por cada nueva posición que ocupa, y se detendrá sólo en las posiciones marcadas por "terminal", en las que incurre en recompensas de 1 ó -1.

El inconveniente de elegir una de las cuatro posibles direcciones (Norte, Sur, Este y Oeste), es que los desplazamientos del agente no son confiables. Es decir, si el agente opta por una dirección específica, su desplazamiento puede darse en esta dirección con probabilidad 0.8, o en dos direcciones adicionales, cada una con probabilidad 0.1. Si la dirección escogida es Norte, el desplazamiento puede efectuase también en dirección Oeste o en dirección Este. Si toma el Este, puede eventualmente desplazarse además hacia el Norte o el Sur. Si elige

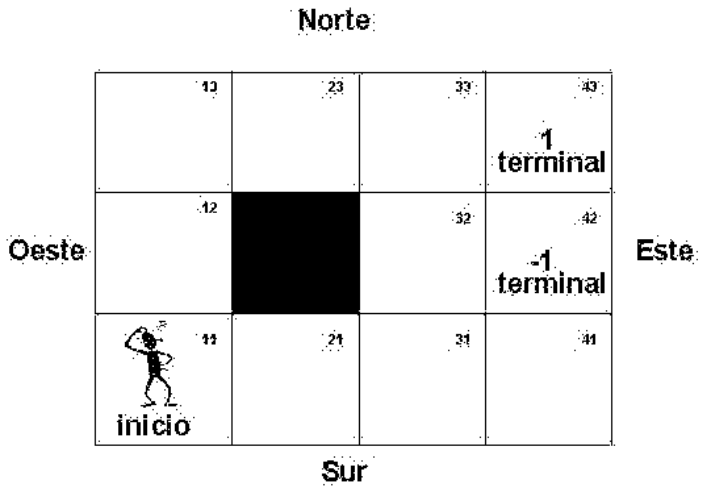

Figura 5.1. Problema de la rejilla 3x2.

la dirección Sur, posiblemente se dirija también en dirección Oeste o Este. Si elige el Oeste puede ir también en dirección Norte o Sur.

Por ejemplo, si el agente estuviese ubicado en la posición (3,2), y decidiera moverse en dirección Norte, llegará a la posición (3,3) con probabilidad 0.8, o permanecerá en la misma posición con probabilidad 0.1 (ya que en dirección Oeste está el objeto y choca contra él), o se moverá a la posición (4,2) con probabilidad 0.1, tal como se representa en la parte (A) de la figura 5.2. Si por el contrario, el agente decide moverse en dirección Este (una elección extremadamente costosa), llegará a la posición (4,2) con probabilidad 0.8, o a la posición (3,3) con probabilidad 0.1 o a la posición (3,1) con probabilidad 0.1, tal como se muestra es la parte (B) de la figura 5.2.

El problema consiste entonces en encontrar la secuencia de direcciones que debe emprender el agente dentro del cuarto que minimicen los costos de su desplazamiento, cuando inicia en la posición (1,1).

El FOMDP asociado a este problema es como sigue:
El conjunto de estados está formado por las 11 posibles posiciones que puede ocupar el agente, es decir

$$X = \{(1,1),(1,2),(1,3),(2,1),(2,3),(3,1),(3,2),(3,3),(4,1),(4,2),(4,3)\}.$$

El conjunto de acciones está formado por las cuatro direcciones en que puede moverse el agente,

$$A = \{Norte, Sur, Este, Oeste\}.$$

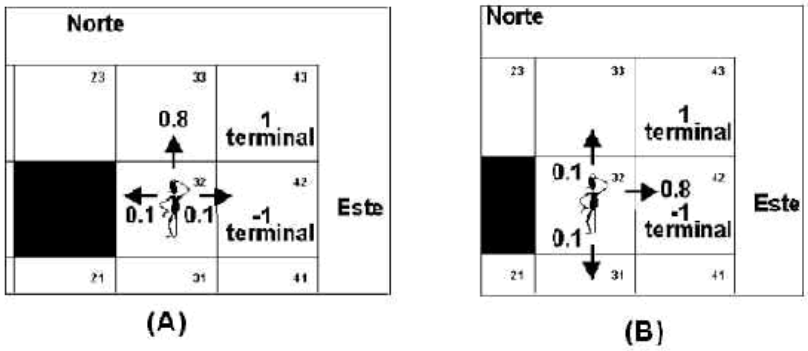

Figura 5.2. Ejemplo de desplazamiento en el problema de la rejilla.

El núcleo estocástico se obtiene al calcular las probabilidades condicionales cuando se fija cada una de las direcciones de desplazamiento.

La función de recompensa está definida como

$$r(x,a) = \begin{cases} 1 & \text{si } x = (4,3) \\ -1 & \text{si } x = (4,2) \\ -0.4 & \text{si } x \neq (4,2) \text{ y } x \neq (4,3), \end{cases}$$

para todo $a \in A$.

El FOMDP ó problema de decisión de Markov se completa al definir como criterio de optimización, la recompensa total esperada con descuento $\beta = 1$.

Utilizando la función recursiva del método de iteración de valores comprobamos que para $n = 20$ y con un error $\varepsilon = 10^{-3}$, se obtiene la política óptima, tal como se muestra en la tabla 5.1. La solución gráfica del problema se representa en la figura 5.3.

6. Conclusiones

El uso de los procesos estocásticos y las herramientas de la programación dinámica estocástica facilitan el estudio teórico de los MDP´s, ya que los resultados de estos tópicos garantizan la existencia de políticas óptimas y se convierten en una herramienta de gran utilidad en la solución de problemas bajo condiciones de incertidumbre en AI.

Los MDP´s presentados aquí, estudiados ampliamente en teoría de las decisiones, no responden a todas las necesidades de la Inteligencia Artificial para resolver problemas de decisión secuencial. Su aplicación en este campo ha motivado numerosas investigaciones recientes, que se alejan del marco teórico de la teoría de la decisión y concentran sus esfuerzos en el desarrollo de algoritmos que disminuyan los costos computacionales.

x	v_{20}	$f^*_{20}(x)$	$\lvert v_{20}(x)-v_{19}(x)\rvert$
(1,1)	0.7053	Norte	3.06038E-05
(1,2)	0.7616	Norte	3.55768E-06
(1,3)	0.8116	Este	1.08269E-06
(2,1)	0.6552	Oeste	8.66968E-05
(2,3)	0.8678	Este	1.80184E-08
(3,1)	0.6112	Oeste	0.000178803
(3,2)	0.6603	Norte	1.16479E-08
(3,3)	0.9178	Este	4.11816E-09
(4,1)	0.3876	Oeste	0.000362292
(4,2)	-1		0
(4,3)	1		0

Tabla 5.1. Resultados en el problema de la rejilla 3x2.

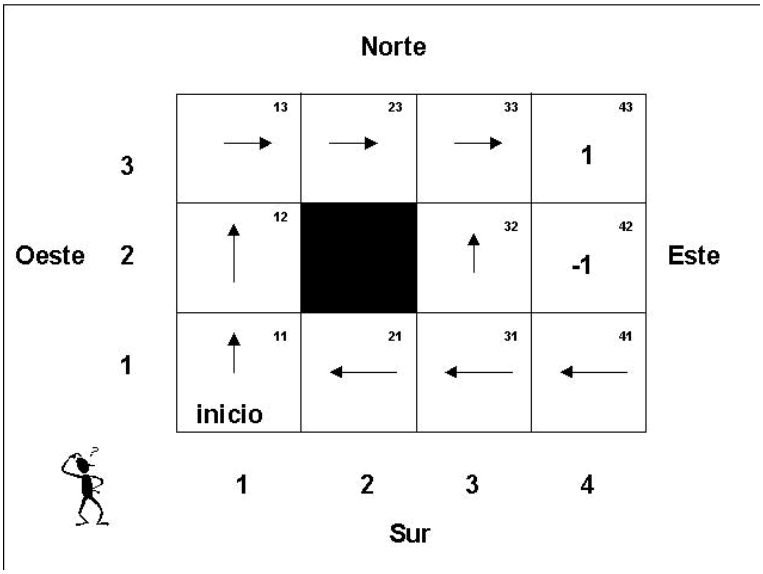

Figura 5.3. Política òptima en el problema de la rejilla 3x2.

La utilización de MDP´s para modelar problemas de planificación bajo condiciones de incertidumbre en el campo de la Inteligencia Artificial, es una alternativa que exige gran cantidad de recursos omputacionales. Para el ejemplo sencillo que hemos presentado, fueron necesarias 20 iteraciones de la función recursiva del método.

El concepto de agente proporciona una mayor claridad en la comprensión de problemas de decisión secuencial en AI y es una herramienta importante en la actualidad para el estudio generalizado de problemas de decisión bajo condiciones de incertidumbre en todas las áreas. La inclusión de teorías rigurosas como la teoría de la decisión y la programación dinámica en la AI, es una de las ganancias más importantes del enfoque basado en agentes.

Bibliografía

[1] R. Bellman, *Dynamic programming*, Princeton Univ. Press, Printeton, New Jersey, Artificial Intelligence Research **11** (1957).

[2] L. Blanco y M. Muñoz, *Martingalas y cadenas de Markov con parámetro de tiempo discreto*, XV Coloquio Distrital de Matemáticas y Estadística, Bogotá, 1999.

[3] L. Blanco y M. Muñoz, *Introducción a la teoría avanzada de la probabilidad*, Universidad Nacional de Colombia, Bogotá, 2002.

[4] C. Boutilier, T. Dean, S. Y. Hanks, *Decision theoretic planning: structural assumption and computational leverage*, Journal of Artificial Intelligence Research **11** (1999).

[5] Y. García, *Procesos Markovianos de decisión en el modelamiento de agentes racionales*, Trabajo de grado, Departamento de Matemáticas y Estadística, Universidad Nacional de Colombia, Bogotá,2002.

[6] O. Hernández Lerma, *Adaptative Markov Control Processes*, Springer Verlag, New York, 1989.

[7] C. J. Himmelberg, T. Parthasarathy, F.S. Van Vleck, *Optimal plans for dynamic programming problems*, Math. Oper. Res. **1** (1976), 390–394.

[8] E. Kreyzig, *Introductory Functional Analysis with Applications*, John Willey & Sons. New York, 1978.

[9] S. Mahadervan *To discount or not discount in reforcement learning: A case study in comparing R-learning and Q-learning*, Proccedings of the of the Eleventh International Conference on Maching Learning, New Brunswick, N. J. (1994), 164–172.

[10] N. Oza, *Markov decision problems*, Departament of computer science, Berkeley University, 1995.

[11] J. M. Puterman, *Markov Decision Processes*, D. P. Heyman and Sobel. Eds., Handbook in Operations Research and Management Science, Vol 2. Stochastic Models, North Holland, 1990.

[12] S. J. Rusell, P. Norving, *Inteligencia Artificial: un enfoque moderno*, Primera Edición en Español, Prentice Hall Hisopanoamericana, 1996.

[13] L. S. Shapley, *Stochastic games*, Proc. Natl. Acad. Sci. USA. **39** (1953), 1095–1100.

[14] J. Van Der Wal, *Stochastic Dynamic Programming*, Trac 139, Mathematical Centre, Amsterdam, 1984.