ROBERTO QUINDE

PROCESOS DE DECISIÓN DE MARCOV

Para contexualizar que es un proceso de desción de marcov conocido como MDP, decimos que existe la posibilidad de que una accion realizada en un estado que produzca una distribucion de probabilidades sobre los efectos posibles. En cada paso el agente recibe una recompensa nemerica que en general depende del estado actual y da la accion aplicada. De tal manera que el objetivo es encontrar una estrategia reactiva de control o politica de accion que maximize la recompensa en el tiempo.

CONCEPTO

Las tecnicas basadas en los Procesos de Decision de Markov (MDP) modelan un problema de decision secuencial, en el que un sistema evoluciona en el tiempo y es controlado por un agente. La dinamica del sistema es gobernada por una funcion de transicion probabilısta que asocia estados y acciones con nuevos estados, y que satisfacen la propiedad de Markov (poder de prediccion del efecto de una accion sobre un estado).

El principio de los algoritmos estandar para MDPs, es resolver el mismo problema de planificacion con incertidumbre que las tecnicas clasicas extendidas. Sin embargo, lo hacen de formas distintas. Por ejemplo, mientras un MDP encuentra polıticas, los planificadores clasicos encuentran secuencias de acciones. La mayorıa de los planificadores clasicos describen el estado de un sistema como un conjunto de clausulas logicas. Los MDPs en general lo hacen a traves de instancias de las variables relevantes del dominio. Los metodos de solucion de un MDP enfatizan el uso de la programacion dinamica, mientras que los metodos de solucion de la planificacion clasica realizan busqueda sobre espacios de estados y de acciones estructuradas.

Formalmente, un MDP *M* es una tupla $M = \langle S, A, \Phi, R \rangle$ Ocaña M., 2005), donde:

- *S* es un conjunto finito de estados del sistema.
- *A* es un conjunto finito de acciones, que se ejecutan en cada estado.
- $\Phi : AxS \rightarrow \Pi(S)$: es la función de transición de estados dada como una distribución de probabilidades y la cual asocia un conjunto de posibles estados resultantes de un conjunto de acciones en el estado actual. La probabilidad de alcanzar un estado *s'* realizando la acción *a* en el estado *s* se escribe $\Phi(a, s, s')$.
- *R: S* x *A* → *R* es una función de recompensa. *R(s, a)* es la recompensa que el sistema recibe si lleva a cabo la acción *a* en el estado *s*. Esta función define la meta que se quiere alcanzar.

Cabe mencionar que se ha adoptado la notación de Ocaña M. (2005), para la definición formal de un MDP. Sin embargo la función de transición y la función de recompensa pueden representarse como $\Phi : SxAxS \rightarrow \Pi(S)$ y $R: SxAxS \rightarrow R$ respectivamente debido a que en ambos

casos el resultado de cada función está asociada a un estado al que se llega con la acción realizada.

Una política para un MDP es una asociación $\pi : S \rightarrow A$ que selecciona una acción por cada estado, es decir, define cómo se comporta el sistema en un determinado estado, y puede verse como un mapeo de los estados a las acciones.

POLITICAS OPTIMAS

Resolver un MDP consiste en encontrar la política óptima, una directiva de control que maximiza la función de valor sobre los estados. Teóricamente, es posible obtener todas las posibles políticas para un MDP y a continuación escoger entre ellas, aquella que maximiza la función de valor. Sin embargo, este método es computacionalmente costoso, puesto que el número de políticas crece exponencialmente con el número de estados. Existen, sin embargo, otros métodos que permiten seleccionar la política óptima aprovechando la propiedad de que ésta será también localmente óptima para cada estado individual.

Howard (1960) demostró que, para el caso más general de horizonte-infinito, existe una política estacionaria π^* que es óptima para cualquier estado inicial del proceso. La función de valor para esta política viene dada por la solución de la ecuación de Bellman (1):

$$V^*(s) = \max_a \left(r(s,a) + \gamma \sum_{s' \in S} p(s' | s,a) \cdot V^*(s') \right), \forall s \in S$$

(1)

El método tradicional para obtener políticas óptimas es la programación dinámica (Bellman 1957, Bertsekas 1995). Dos de los métodos más utilizados son:

- **El metodo de Iteración de Valor**
- **El de Iteración de Política.**

Los dos metodos se basan en modificar las utilidades de los estados vecinos de manera que satisfagan las ecuaciones de Bellman. La repetición de este proceso de modificación local en cada uno de los estados durante un número suficiente de iteraciones hace converger las utilidades de los estados individuales hacia sus valores correctos. Estos métodos permiten obtener la política óptima fuera de línea, esto es, ayudan a que el agente no invierta tiempo en aprender la política probándola en tiempo de ejecución.

Los algoritmos 1 y 2 resumen respectivamente, los métodos de iteración de política e iteración de valor. El primero consiste en calcular la utilidad de cada uno de los estados y con base en estas utilidades, seleccionar una acción óptima para cada uno de ellos. El segundo funciona escogiendo una política, y luego calculando la utilidad de cada estado con base en dicha política. Luego actualiza la política correspondiente a cada estado utilizando las utilidades de los estados sucesores, lo que se repite hasta que se estabiliza la política.

REFERENCIAS

Ruiz, S., & Hernández, B. (2014). Procesos de decisión de Markov y microescenarios para navegación y evasión de colisiones para multitudes. *Res. Comput. Sci.*, *74*, 103-116.

Reyes Ballesteros, A. (2006). Representación y aprendizaje de procesos de decisión de markov cualitativas.

Siembro, G. C. (2007). Procesos de decisión de Markov aplicados a la locomoción de robots hexápodos. *Recuperado de http://inaoe. repositorioinstitucional. mx/jspui/handle/1009/588*.

Guillén, M. E. L. (2004). *Sistema de navegacion global basado en procesos de decisión de markov parcialmente observables. aplicación a un robot de asistencia personal* (Doctoral dissertation, Universidad de Alcalá).

H. Cruz-Su·rez, Procesos de decisiÛn de Markov descontados: solu- ciones Ûptimas mediante problemas de control determinista dife- renciables, Revista Iberoamericana de Sistemas CibernÈtica e In- form·tica, V. 2 N. 1, 2005. (Disponible en la p·gina web: http://www.iiisci.org/journal/risci/.)