Aprendizaje de selección de acciones en un mundo simple pero impredecible

Sergio A. Rojas[1]
José J. Martínez[2]

Resumen

Uno de los principales problemas estudiados en la simulación de agentes artificiales autónomos es el de la selección de acciones: un mecanismo que le permita al sistema escoger la acción más apropiada para la situación en que se encuentre, de tal forma que maximice su medida de éxito. El aprendizaje por refuerzo representa un enfoque atractivo para atacar este problema, ya que se basa en la búsqueda de señales de premio y la evasión de señales de castigo mediante un proceso de ensayo y error. En este artículo presentamos al PAISA I, una criatura artificial que aprende a comportarse (seleccionar acciones) utilizando una técnica de aprendizaje por refuerzo (aprendizaje Q) para optimizar la cantidad de comida que puede encontrar en un mundo impredecible, aunque con un espacio estado-acción pequeño.

Palabras clave:
aprendizaje por refuerzo, aprendizaje Q, agentes autónomos, animats.

Abstract

One of the main problems studied in simulation of artificial autonomous agents is the action-selection: a mechanism that allows the system to choice the more suitable action for the specific situation where it is located, in such a way that maximises his success measure. The reinforcement learning represents an attractive approach to attack this problem, cause it is based in the searching of awards signals and the refusing of punishments by a trial and error process. In this paper, we present the PAISA I, an artificial creature that learns to behave (that is, action-selection) using a reinforcement learning technique known as Q-learning, to optimise the amount of food that he can find in an unpredictable world, although in a small state-action space.

Keywords:
Reinforcement learning, Q-learning, autonomous agents, animats.

1 Grupo de Interés Adaptación Computación & Mente
Facultad de Ingeniería
Universidad Distrital Francisco José de Caldas
Santafé de Bogotá, Colombia
e-mail: sar@lettera.net

2 Departamento de Ingeniería de Sistemas
Universidad Nacional de Colombia
Ciudad Universitaria
Santafé de Bogotá, Colombia
e-mail: josej@ing.unal.edu.co

Introducción

La construcción de sistemas autónomos que puedan desempeñarse de manera independiente en el ambiente en el que se encuentran situados (es decir, en la tarea que deben solucionar), enfrentando situaciones inesperadas para las cuales nunca habían sido programados y que a la postre deben resolver exitosamente para seguir funcionando eficazmente (esto es, sobreviviendo), es uno de los temas que mayor interés ha suscitado dentro del área de estudio de la inteligencia artificial (IA).

Particular importancia han tomado en los últimos años los estudios sobre simulación del comportamiento adaptativo que poseen los animales, convirtiéndose en un nuevo enfoque para la creación de este tipo de sistemas. Este nuevo paradigma se ha denominado enfoque "animat" (del inglés artificial animal), y se fundamenta en especial en la utilización de conceptos aplicados al estudio de animales reales, tomados principalmente de la biología y complementados con los avances alcanzados en la teoría de sistemas computacionales complejos, con el ánimo de convertir al computador en una potente herramienta útil para la simulación de agentes autónomos.

En este artículo presentamos al Prototipo Animat de Interacción Simple con el Ambiente I (PAISA I) [Rojas, 1998], un agente autónomo que aprende a desenvolverse en un mundo artificial poblado de comida, construyendo una política adecuada de selección de acciones para maximizar el premio recibido durante su experiencia. En la sección 1 se revisa el fundamento teórico utilizado para su construcción; en la sección 2 se explica más detalladamente el experimento y su implementación; en la sección 3 se presentan los resultados obtenidos, y, para finalizar, la sección 4 está dedicada a las conclusiones.

1. Aprendizaje por refuerzo

Una de las cuestiones fundamentales que se deben tener presentes en la construcción de agentes autónomos es la política o estrategia de selección de acciones que permita al sistema enfrentar las situaciones amenazantes que se le presenten durante su actividad. La teoría clásica de IA se basa en dotar al sistema con una representación simbólica del conocimiento necesario para resolver satisfactoriamente tales situaciones. El problema se complica cuando éstas aparecen inesperadamente; en tal caso, el sistema debería ser capaz de adquirir nuevo conocimiento, para modificar su representación interna y "saber" enfrentar exitosamente la situación cuando se le presente de nuevo.

Este aprendizaje es una de las debilidades de la teoría clásica, y para muchas aplicaciones (especialmente problemas de clasificación y reconocimiento), ha sido superado mediante un enfoque de aprendizaje supervisado, en el cual existe un tutor externo, que mediante ejemplos le enseña al sistema a controlar la situación asociando parejas estimulo-respuesta. El entrenamiento se demora hasta que el agente demuestra que ha aprendido tanto como lo que el tutor sabe. Infortunadamente, en muchos casos de la vida real no es posible recolectar de antemano el conocimiento necesario (tutor) ni tampoco se dispone del tiempo suficiente para un entrenamiento fuera de línea.

El aprendizaje por refuerzo [Sutton et al., 1998] representa un enfoque alternativo para resolver este inconveniente. Se fundamente en el hecho de que, mediante un proceso de ensayo y error, el agente puede ser capaz de descubrir las acciones buenas y malas, al recibir retroalimentación constante del ambiente en forma de señales de premios o castigos. De esta manera va adquiriendo conocimiento en línea, construyendo una política de selección de acciones que le permite ser exitoso dentro del ambiente que lo rodea, y además, le permite modificar su comportamiento ante cambios amenazantes. La gran ventaja de este tipo de aprendizaje en el cual es el ambiente el que envía una señal de aceptación hacia el sistema, es su posibilidad de aplicación en muchos campos de la vida real (podíamos nombrar, entre otros, navegación, control, asignación, programación, búsqueda, evasión).

Uno de los métodos de aprendizaje por refuerzo que ha ganado gran interés en los últimos años es el aprendizaje Q[3] [Watkins et al., 1992], basado en un algoritmo incremental sencillo desarrollado a partir de la teoría de la programación dinámica para el aprendizaje postergado [Peng, 1993]. La idea principal de este algoritmo es mantener una valoración de la calidad (valor Q) para cada posible asociación (x, a) del espacio estado-acción del agente. Este valor Q representa un estimado del refuerzo total esperado postergado que recibirá el agente en el largo plazo por ejecutar la acción a al encontrarse en el estado x. La variación incremental de los valores Q durante dos pasos consecutivos en el tiempo garantiza la construcción de una función evaluadora óptima que permita la utilización de una política confiable de selección de acciones[4] . Así, los valores Q son ajustados de acuerdo con la ecuación (1),

$$Q_{n+1}(x,a) = (1-\alpha)Q_n(x,a) + \alpha(r+\gamma V^*(y)) \quad (1)$$

donde:

3 Traducción libre del término en inglés Q-learning.

4 Watkins et al., 1992 han demostrado la convergencia del algoritmo hacia una política óptima con una probabilidad de 1.

x: Estado actual percibido por el agente

a: Acción tomada por el agente

y: Siguiente estado percibido por el agente, después de ejecutar a

$Q(x,a)$: Calidad de ejecutar a, estando en x, y a continuación seguir con la política óptima de selección de acciones

$V^*(x) = max_a Q(x,a)$

r: Refuerzo inmediato recibido por el agente

$0<\gamma<1$: Factor de relevancia de los refuerzos anteriormente recibidos

$0<\alpha<1$: Factor de ponderación de la calidad

Evaluando la función de calidad Q, el agente puede encontrar la acción con una mayor valía, la cual, intuitivamente, es la más apropiada para ejecutar en el estado actual.

2. El animat PAISA I

PAISA I [Rojas, 1998] es un animat motivado a conseguir la comida que encuentre en su mundo a medida que se mueve por él. Su meta es gastar la menor cantidad de energía mientras busca la comida. Su mundo es discreto y consta de una cuadrícula toroidal dentro de la cual cada celda puede contener una partícula de comida o estar vacía (véase Figura 1).

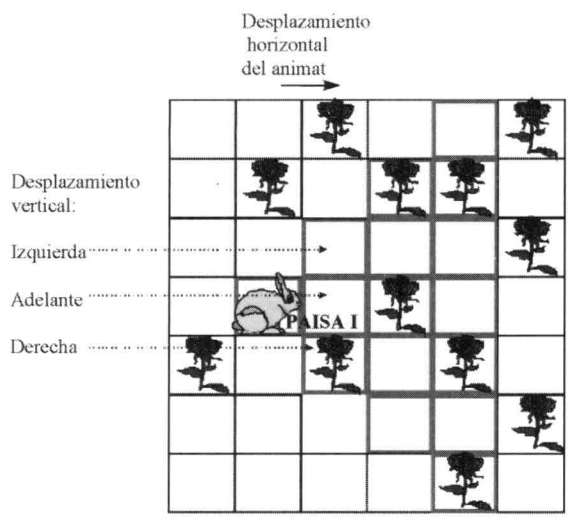

Figura 1. El PAISA I y su ambiente.

El repertorio de acciones de PAISA I es limitado: siempre se mueve hacia adelante en dirección izquierda, enfrente o derecha. Cada movimiento le representa un gasto de energía y con él puede adquirir máximo una partícula de comida (en forma de energía). La comida crece aleatoriamente. Su tarea consiste en seguir un camino que le reporte el mayor beneficio en adquisición de energía, es decir, debe minimizar el número de movimientos a celdas vacías. PAISA I recibe información de su mundo mediante un campo de visión de forma cónica (mostrado con líneas más gruesas en la Figura 1). La profundidad de este campo es un factor importante para la planeación que pueda hacer PAISA I, ya que no sólo debe detenerse en la acción que le represente un premio inmediato, sino también en aquella que en el largo plazo será más beneficiosa (por ejemplo, al bajar una celda en el tiempo n, será imposible que alcance en n + 1 una casilla que estaba dos celdas arriba en el tiempo n). Por esta razón, aunque el espacio estado-acción del agente es relativamente pequeño, el mundo es suficientemente impredecible como para realizar un experimento de aprendizaje por refuerzo.

El mecanismo de selección de acciones utilizado por PAISA I se basó en el algoritmo de aprendizaje Q mediante una implementación tabular. La simulación se realizó con el lenguaje de programación C, en una máquina Pentium de 233Mhz.

3. Resultados del experimento

PAISA I demostró una adaptación exitosa a su ambiente impredecible, al aprender a solucionar el problema en el que estaba situado, que para este caso fue obtener la mayor cantidad de comida para maximizar su energía. En configuraciones de comida particulares, donde el crecimiento no era aleatorio sino un patrón establecido, PAISA I fue capaz de encontrar la regularidad observada y aprender a explotarla sin desperdiciar energía en movimientos erróneos. En la Figura 2 se observa una secuencia de los pasos representativos de una simulación del comportamiento seguido con una configuración de tipo sinusoidal.

Estas adaptaciones se alcanzaron rápidamente en pocos pasos de simulación. Adicionalmente, cuando ocurrieron cambios inesperados, por ejemplo cuando se desplazaba el patrón de crecimiento, PAISA I reaccionó un poco desconcertado pero después de un tiempo (corto igualmente), logró readaptarse al ambiente.

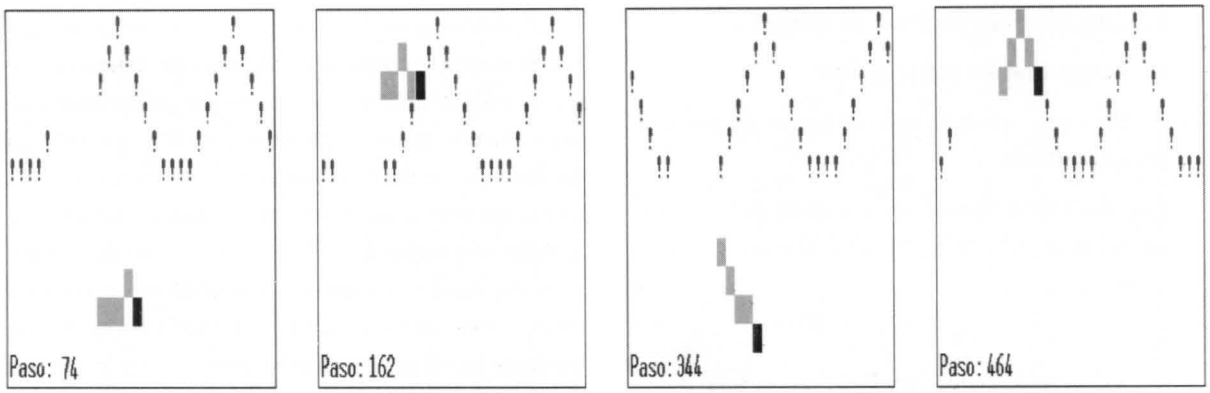

Figura 2. PAISA I navegando por un mundo con un patrón de comida regular.

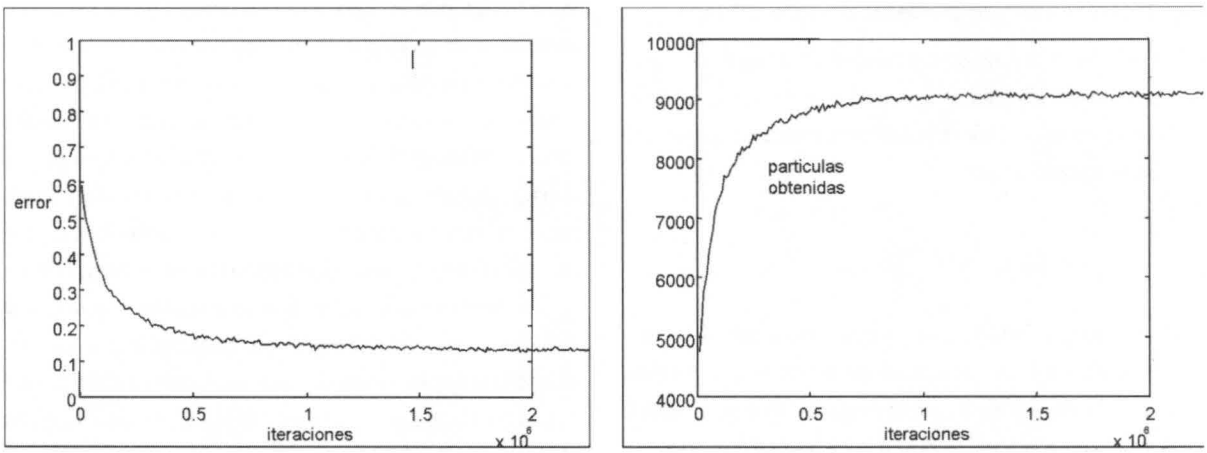

Figura 3. Gráficas de la tasa de error del comportamiento PAISA I y del número de partículas de comida obtenidas.

Por otra parte, realizamos experimentos en mundos con crecimiento aleatorio de comida, con el ánimo de estimar una medida de éxito del comportamiento de PAISA I. Para ello, se implementó una función de error basada en una medida de probabilidad condicionada de obtener comida dado un estado determinado por el campo de visión. Una gráfica representativa de la ejecución del experimento se puede observar en la Figura 3. En ella se encuentra que el comportamiento de PAISA I mejora (su tasa de error disminuye), mientras más experiencia tiene con el ambiente. De manera análoga, en la misma figura se puede ver cómo aumenta la obtención de partículas de comida en función del tiempo.

Durante las simulaciones encontramos algunos aspectos interesantes sobre la influencia de los parámetros del modelo. Por ejemplo, observamos que la cantidad de comida disponible en el mundo es un factor importante para el desempeño de PAISA I: a mayor comida disponible, menor esfuerzo hace por aprender, ya que a pesar de ejecutar acciones calificadas como erróneas, podía corregirlas fácilmente gracias a la abundancia de premios. Otro factor significativo fue el módulo de exploración que se le acondicionó, gracias al cual, durante la parte inicial del experimento, se comportaba más por ensayo y error que confiado en la política dada por el aprendizaje Q, mientras que al final sí explotaba la estrategia aprendida; esto le permitía buscar por todo el ambiente las condiciones más favorables para su comportamiento; muchas veces al eliminar este módulo, PAISA I se aferraba a mínimos locales sin encontrar la solución óptima. Es esencial también resaltar la influencia del gasto de energía en cada paso de simulación y la amplitud de su campo receptor: el primero lo incentiva a disminuir su tasa de error más rápido, mientras que el segundo lo estimula a tener una mayor capacidad de planeación.

4. Trabajo futuro

PAISA I obtuvo buenos resultados y demostró capacidad de adaptación a un mundo dinámico e impredecible utilizando un método de aprendizaje por refuerzo. Sin embargo, aunque el experimento trató elementos complejos (como ambiente variable, adaptación, exploración-explotación), se desarrolló con un espacio estado-acción pequeño, característica no muy común en problemas cercanos a la realidad. El trabajo que sigue a continuación es complicar un poco tanto el repertorio de acciones como el ambiente que rodea al PAISA: mundo continuo, movimiento en todas las direcciones, ubicación de otros agentes (como depredadores o parásitos). El aprendizaje Q parece ser una técnica promisoria en el área del aprendizaje por refuerzo, tal como se demostró en este experimento; sin embargo, cuando el espacio estado-acción crece, su implementación tabular ya no es factible, debido a problemas de memoria y tiempo de convergencia. Una posible solución, que nos parece atractiva, es aproximar la función Q mediante una red neuronal artificial que sea capaz de generalizar para no tener que almacenar la calidad de todas las posibles combinaciones estado-acción sino de un subconjunto representativo. La red puede entrenarse con el algoritmo de retropropagación tomando la salida deseada como la parte derecha de la ecuación (1) y la salida actual la producida por la red al presentarle el estado x (véase Lin, 1992). Este tema es de gran interés actualmente, y en torno a él se han presentado otros estudios como el aprendizaje Q particionado [Munos et al., 1994], enfoques evolutivos con adaptación simbiótica [Moriarty et al., 1996], y otros avances en algoritmos de aprendizaje por refuerzo como Q multipaso [Peng et al., 1996] o actualización de la ventaja [Baird, 1994].

Así pues, nuestra investigación seguirá enfocada en este sentido, pues representa un desafío interesante para la obtención de resultados que más adelante puedan ser utilizados en aplicaciones ingenieriles reales. PAISA I fue un experimento agradable pero sobre todo estimulante para continuar con una amplia gama de proyectos sobre animats y comportamiento adaptativo.

Agradecimientos

Parte de esta investigación fue patrocinada por el Programa de Cooperación Interuniversitaria de la Agencia Española de Cooperación Internacional.

Deseamos expresar nuestro agradecimiento especial al doctor Francisco Vico, del Departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga, por su contribución en este trabajo.

Bibliografía

[Baird, 1994] Baird, L. C. (1994). Reinforcement Learning in Continuous Time: Advantage Updating. Proceedings of the International Conference on Neural Networks.

[Lin, 1992] Lin, L. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. En: Machine Learning, 8.

[Moriarty et al., 1996] Moriarty, D. E.; Miikkulainen, R. (1996). Efficient reinforcement learning through symbiotic evolution. En: Machine Learning, 22.

[Munos et al., 1994] Munos, R.; Patinel, J. (1994). Reinforcement learning with dynamic covering of state-action: partitioning Q-learning. En: Cliff, D.; Husbands, P.; Meyer, J. A.; Wilson, S. W. (Eds), From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior. The MIT Press/Bradford Books.

[Peng, 1993] Peng, J. (1993). Efficient Dynamic Programming-based Learning for Control. Tesis doctoral. College of Computer Science of Northeastern University.

[Peng et al., 1996] Peng, J.; Williams, R. J. (1996). Incremental Multi-step Q-Learning. En: Machine Learning, 22.

[Rojas, 1998] Rojas, S. A. (1998). Disertación teórica sobre simulaciones inspiradas biológicamente para el estudio del comportamiento adaptativo. Monografía de grado. Facultad de Ingeniería de la Universidad Nacional de Colombia.

[Sutton et al., 1998] Sutton, R. S.; Barto, A. G. (1998). Reinforcement Learning: An Introduction. The MIT Press.

[Watkins et al., 1992] Watkins, C. J.; Dayan, P. (1992). Q-Learning. En: Machine Learning, 8.