Prototipo Animat de interacción simple con el ambiente I: un experimento de aprendizaje maquinal

Sergio A. Rojas[1]

RESUMEN

Los agentes artificiales autónomos deben demostrar características de aprendizaje que les permita comportarse exitosamente ante eventos inesperados para los cuales no hayan sido programados. A medida que el agente experimente con su mundo, debe descubrir información útil, representarla y adquirirla para posteriormente utilizarla de tal forma que maximice su medida de éxito cuando se enfrente a situaciones similares. Una de las técnicas más interesantes que se ha trabajado recientemente es el aprendizaje por refuerzo, el cual 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 se describe el agente autónomo PAISA I, un animal artificial (animat) que aprende una política adecuada de selección de acciones para maximizar 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.

Autonomous artificial agents should have learning features that allow them to behave successfully in the presence of unexpected events which they are not programmed for. The more experience the agent have with its world, the more useful information it can acquire for enhance its behavior in front of similar situations. A recent and interesting technique is reinforcement learning which is based in the agent seeking for rewards and avoiding for punishments like a trial-and-error approach. This paper describes the PAISA I an autonomous agent that can learn an adequate action-selection policy for maximize the amount of food it can find in a unpredictable world, in spite of a small state-action space.

Key words: reinforcement learning, Q-learning, autonomous agents, animats.

El PAISA es un animal artificial simulado (animat) que aprende con éxito a buscar comida en un mundo desconocido para él.

I. INTRODUCCIÓN

La construcción de sistemas que puedan desempeñarse de manera autónoma e independiente en el ambiente en el que se encuentran situados es uno de los temas que mayor interés ha tomado dentro del área de estudio de la inteligencia artificial (IA). Los agentes deben resolver la tarea asignada, y del mismo modo enfrentar situaciones inesperadas para las cuales nunca habían sido programados y que a la postre deben resolver exitosamente para seguir funcionando de manera eficaz, esto es, para sobrevivir [1].

Una de las áreas de mayor actividad académica de los últimos años es el estudio sobre simulación del comportamiento adaptativo que poseen los animales, como un nuevo enfoque para la creación de agentes autónomos. Los *animats,* como se denomina a este tipo de sistemas (del inglés *artificial animal*), se fundamentan en conceptos aplicados al estudio de animales reales, tomados principalmente de la etología y complementados con avances alcanzados en la teoría de sistemas computacionales complejos.

A continuación se describe un primer experimento desarrollado con el Prototipo Animat de Interacción Simple con el Ambiente (PAISA I), un agente autónomo que aprende a comportarse en un mundo artificial, utilizando una política de selección de acciones que construye mediante un aprendizaje por refuerzo. El artículo se encuentra organizado como sigue: En la sección I se revisa el fundamento teórico del animat; en la sección II se explica con detalle el experimento realizado; en la sección III se presentan algunos resultados; al finalizar se plantean las conclusiones y se discuten ideas para un trabajo futuro.

I. MÁQUINAS QUE APRENDEN POR PREMIOS Y CASTIGOS

Una de los aspectos fundamentales de los agentes autónomos, es la política o estrategia de selección de acciones, pues es la que define el comportamiento que tiene el agente ante situaciones amenazantes. La teoría clásica de IA utiliza una base de conocimientos a partir de la cual, utilizando reglas de producción se pueden hacer inferencias que produzcan respuestas a la situación inesperada [2]; el inconveniente radica en que si el agente no tiene suficiente información, puede verse impedido para generar tal respuesta. Una alternativa que se ha propuesto a este problema es el uso de redes neuronales artificiales, que han demostrado características como la generalización y la tolerancia al ruido; sin embargo, usualmente estos sistemas requieren de un entrenamiento supervisado que se realiza fuera de línea y que puede llegar a consumir un tiempo considerable. Desafortunadamente, los agentes autónomos tienen que

[1] Director del Grupo de Interés en Adaptación, Computación & MEnte (ACME- Universidad Distrital Francisco José de Caldas).

Artículo recibido en Septiembre 10 de 2002
Aceptado el 13 de Noviembre 2002

Este animat utiliza una técnica denominada Aprendizaje Q mediante la cual construye una política de selección de acciones óptimas.

experimentar con el ambiente y aprender en tiempo real, por lo que muchas veces no es posible recolectar de antemano el conocimiento necesario (tutor o base de conocimientos), o no se dispone del tiempo suficiente para un entrenamiento.

El aprendizaje por refuerzo [3] aparece como una propuesta interesante para el desarrollo de agentes autónomos. Se fundamente en un proceso de ensayo y error, mediante el cual el agente puede descubrir las acciones buenas y malas al recibir retroalimentación constante del ambiente en forma de señales de premios o castigos[1]. Así, puede ir construyendo una política de selección de acciones que le permita ser exitoso dentro de su ambiente, además de poder modificar su comportamiento ante cambios amenazantes, convirtiéndolo en un agente adaptativo.

Uno de los métodos de aprendizaje por refuerzo que ha ganado gran interés en los últimos años es el aprendizaje Q [2], un algoritmo incremental desarrollado a partir de la teoría de la programación dinámica para el aprendizaje postergado [4, 5]. La idea principal consiste en almacenar una valoración de la calidad (valor Q) para cada posible asociación (x, a) del espacio estado-acción del agente. Tal valor Q, representa un estimado del refuerzo total esperado 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[3]. Los valores Q son ajustados de acuerdo a la ecuación (1),

$$Q_{n+1}(x,a) = (1-a)Q_n(x,a) + a\left(r + yV^0(y)\right) \quad (1)$$

en donde:

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
$V^0(x) = max_Q\, Q(x,a)$	El máximo valor de calidad para todas las posibles acciones que pueda ejecutar el agente estando en x
r	Refuerzo inmediato recibido por el agente (premio o castigo)
$0 < \gamma < 1$	Factor de relevancia de los refuerzos anteriormente recibidos
$0 < \alpha < 1$	Factor de ponderación del valor de calidad

Al evaluar la función de calidad Q, el agente puede seleccionar la acción con una mayor valía, que será intuitivamente la mejor para ejecutar en el estado en que se encuentre. Obsérvese que una vez ejecutada una acción, el agente debe actualizar el valor Q para la pareja *(x, a)* escogida, utilizando el valor *r* del refuerzo enviado por el ambiente. Si el refuerzo es positivo, su valor de calidad aumentará, de lo contrario, disminuirá.

II. DESCRIPCIÓN DEL PAISA I

PAISA I [6] es un animat que tiene como tarea conseguir comida en su mundo mientras se mueve por él. Aunque se trata de un agente muy simple, el objetivo principal de su desarrollo fue comprobar que podía aprender un comportamiento desde cero, utilizando el aprendizaje Q. Su meta es gastar la menor cantidad de energía mientras busca la comida en una cuadrícula discreta toroidal. El animat se mueve entre celdas que pueden estar vacías o contener partículas de comida (Fig. 1).

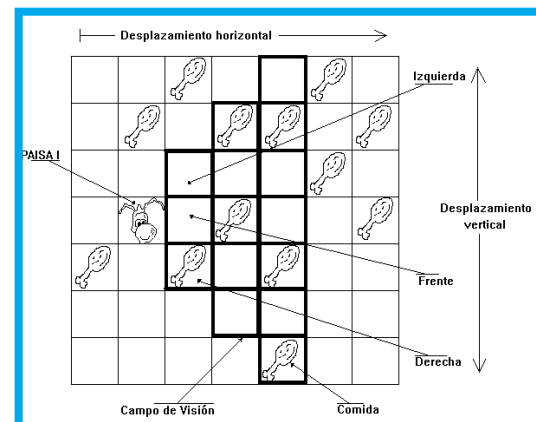


Figura 1. El animat PAISA I en su mundo

El repertorio de acciones de PAISA I es limitado: siempre se mueve hacia adelante en dirección izquierda, enfrente o derecha. En cada movimiento gasta una unidad de energía (castigo) y puede adquirir máximo una partícula de comida (premio). Para detectar la comida, PAISA I tiene un campo de visión de forma cónica (mostrado con líneas más gruesas en la Fig.1); la profundidad de este campo es un factor importante para que el comportamiento no sea simplemente reaccionario (para el paso inmediato) sino planificador (descubriendo rutas con mayor concentración de comida que pueda alcanzar en pasos posteriores).

La comida del mundo puede crecer aleatoriamente o con un patrón determinado. Se busca que el agente construya una política de selección de acciones para explorar una ruta que le reporte el mayor beneficio en adquisición de energía, es decir, que trate de pisar siempre celdas con comida y evite pasar por celdas vacías. De esta forma, aunque las características del animat son limitadas (espacio estado-acción pequeño), las condiciones del ambiente son lo suficientemente impredecibles como para que el agente demuestre un aprendizaje exhibiendo un comportamiento adaptativo.

1 Este tipo de aprendizaje es el que se utiliza para entrenar los animales. Cualquiera que tenga un perro en su casa puede intentar enseñarle a dar la mano ofreciéndole una galleta cada vez que lo haga bien, o despreciándole cada vez que lo haga mal. Se basa en el condicionamiento clásico descubierto por Pavlov a principios del siglo pasado.

2 Traducción libre del término en inglés *Q-learning*.

3 Watkins [4] ha demostrado la convergencia del algoritmo hacia una política óptima con probabilidad de 1.

En diferentes experimentos el agente aprendió a ser confiado cuando veía abundancia de comida, y precavido cuando recibía castigos más fuertes.

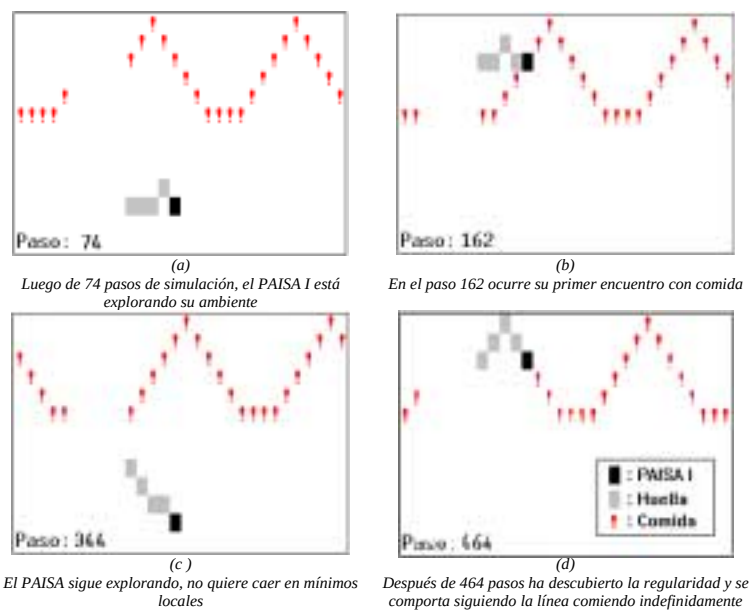
El aprendizaje del PAISA I, se realizó utilizando la técnica de aprendizaje Q, con una implementación tabular (se mantiene una tabla para almacenar el valor de calidad del espacio estado-acción). La simulación se realizó con el lenguaje de programación C, en una máquina Pentium de 233Mhz.

III. RESULTADOS: COMPORTAMIENTO DEL PAISA I

Los experimentos realizados con el agente PAISA I demostraron su capacidad para adaptarse de manera exitosa. Lo más importante es que su comportamiento fue totalmente aprendido, es decir, una vez ubicado, al PAISA I no se le suministró información de antemano acerca de su ambiente impredecible; simplemente se le premiaba cada vez que seguía rutas donde conseguía comida continuamente, y se le castigaba cuando vagaba por caminos desiertos malgastando su energía.

En configuraciones de comida particulares, donde el crecimiento estaba determinado por un patrón establecido, el animat fue capaz de descubrir la regularidad observada y aprender a explotarla sin desperdiciar energía en movimientos erróneos. En la Fig. 2 se presenta un ejemplo representativo de uno de los comportamientos aprendidos. Allí se observa un mundo con un crecimiento de comida en forma de onda sinusoidal. El PAISA I inicialmente exploró el ambiente cruzando por regiones desiertas, para después de un tiempo rastrear la línea de comida sin salirse de su camino.

Las adaptaciones del PAISA I (utilizando siempre el mismo algoritmo Q para construir su política de selección de acciones) se alcanzaron rápidamente en pocos pasos de simulación. Se probaron diferentes ambientes. En patrones de comida aleatorios, el agente escogía como mejor acción no caminar por celdas vacías, y además planificaba para escoger rutas que le garantizaban comer partículas durante varios pasos de tiempo seguidos. En otros experimentos, el crecimiento de comida se restringió a solo algunas regiones del mundo, y efectivamente el PAISA I decidió moverse siempre dentro de la región sin escapar de ella.



(a) *Luego de 74 pasos de simulación, el PAISA I está explorando su ambiente*

(b) *En el paso 162 ocurre su primer encuentro con comida*

(c) *El PAISA sigue explorando, no quiere caer en mínimos locales*

(d) *Después de 464 pasos ha descubierto la regularidad y se comporta siguiendo la línea comiendo indefinidamente*

Figura 2. Comportamiento del PAISA I con un patrón de comida sinusoidal

Adicionalmente, uno de los resultados más interesantes se encontró cuando al PAISA I ya adaptado a un patrón de comida regular, se le cambiaba repentinamente las condiciones. En este caso, su comportamiento se re-adaptaba rápidamente en menos de diez pasos de simulación a la nueva situación, demostrando que su aprendizaje no solo era óptimo, sino que podía modificarse en tiempo real.

Con el fin de establecer una medida de error del comportamiento del animat, se implementó una función de error basada en la probabilidad condicionada de obtener comida, dado un estado determinado por la información captada en el campo de visión. Luego se comparó el número de partículas de comida que obtendría utilizando esta política ideal con la política aprendida por el animat; los resultados se muestran en la Fig. 3. Allí se puede observar que la tasa de error disminuye a medida que aumenta el tiempo de experimentación con el ambiente. De manera similar, la obtención de partículas de comida aumenta en función del tiempo (mientras más aprende, mejor se comporta).

Durante las simulaciones se encontraron algunos aspectos interesantes sobre la influencia de los parámetros del modelo. Por ejemplo la cantidad de comida disponible en el mundo, es un factor importante para el desempeño de PAISA I: a mayor comida, menor es el esfuerzo por aprender, ya que a pesar de cometer muchos errores, fácilmente consigue premios gracias a la abundancia de comida. Otros parámetros influyentes fueron el gasto de energía que implicaba moverse y la amplitud de su campo de visión: el primero lo incentiva a disminuir su tasa de error más rápido en proporción a un mayor gasto (castigo más fuerte); el segundo lo estimula a tener una mayor capacidad de planeación (o no ser inmediatista).

En algunos experimentos, el PAISA I se aferraba a mínimos locales que le impedían la construcción de una política óptima de selección de acciones. Con el fin de evitarlos se le añadió un módulo de exploración mediante el cual, durante la parte inicial del experimento se comportaba más aleatoriamente (por ensayo y error) que confiado en la política dada por el aprendizaje Q, mientras que al final, si explotaba la estrategia aprendida; esto le permite buscar por todo el ambiente las condiciones más favorables para su comportamiento.

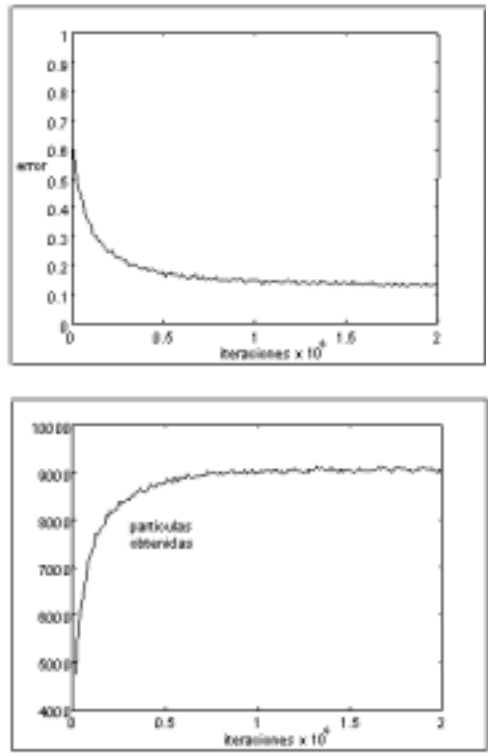


Figura 2. Comportamiento del PAISA I con un patrón de comida sinusoidal

VI. CONCLUSIONES Y PERSPECTIVAS

PAISA I demostró capacidad de adaptación a un mundo impredecible utilizando un método de aprendizaje por refuerzo. Aunque el experimento incluyó elementos complejos (como ambiente dinámico, adaptación, exploración-explotación), hay que resaltar el hecho de que el agente contemplaba un espacio estado-acción pequeño, característica no muy común en problemas cercanos a la realidad. El experimento puede mejorarse escalando tanto el repertorio de acciones como el ambiente que rodea al PAISA: movimientos continuos en todas las direcciones, interacción con otros agentes que representen amenazas (como depredadores o parásitos), utilización de motivaciones internas (como sentir hambre o sed).

Aunque en este experimento se demostró que el algoritmo Q es una técnica eficaz para el aprendizaje en línea, cuando el espacio estado-acción crece, presenta complicaciones en su implementación tabular debido a problemas de memoria y tiempo de convergencia. Una posible solución para este problema sería aproximar la función Q mediante una red neuronal artificial que generalice y por lo tanto almacene un conjunto representativo del valor de calidad de todas las posibles combinaciones estado-acción. La red podría entrenarse con el algoritmo de retropropagación tomando como salida deseada la parte derecha de la ecuación (1) y la salida actual la producida por la red al presentarle el estado x (ver [7]). Este problema de escalamiento del algoritmo para espacios más complejos es de gran interés actualmente y en torno a él se han presentado otros estudios como el aprendizaje Q particionado [8], enfoques evolutivos con adaptación simbiótica [9], y otros avances en algoritmos de aprendizaje por refuerzo como Q multipaso [10] o actualización de la ventaja [11]. Estos serán algunos de los aspectos que se tratarán en futuros trabajos. El PAISA I no solo fue un experimento agradable sino además estimulante para continuar con una amplia gama de proyectos sobre animats y comportamiento adaptativo.

El PAISA I fue un experimento estimulante que incita a profundizar en el estudio sobre animats y comportamiento adaptativo.

VII. AGRADECIMIENTOS

Esta investigación fue parcialmente patrocinada por el Programa de Cooperación Interuniversitaria de la Agencia Española de Cooperación Internacional. Nuestro agradecimiento especial par el Dr. Francisco Vico del Departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga, por su contribución en este trabajo.

REFERENCIAS

[1] Rojas, S.A. et al. (2001). "Supervivencia emergente en un ecosistema presa-depredador artificial". En: *Memorias de la Primera Conferencia Iberoaméricana de Matemática Computacional*. Thomson Learning.

[2] Rusell, S.; Norvig, P. (1996). *Inteligencia Artificial. Un enfoque moderno*. Prentice Hall.

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

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

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

[6] 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.

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

[8] 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.

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

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

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

Autor: Sergio A. Rojas
Ing. de Sistemas, U. Nacional de Colombia. Especialista en Ing. de Software, U. Distrital. Profesor/investigador de la Facultad de Ingeniería, Universidad Distrital FJC. Miembro del Laboratorio de Automatización, Microelectrónica e Inteligencia Computacional (LAMIC). **srojas@udistrital.edu.co**