# Comparación de Algoritmos de Aprendizaje por Refuerzo en la Navegación de Robots Móviles: Q-Learning, Montecarlo y SARSA

David Fuentelsaz Rodríguez
*Dpto. Ciencias de la Computación e Inteligencia Artificial*
*Universidad de Sevilla*
Sevilla, España
davfuerod@alum.us.es

Nombre y apellidos alumno 2
*Dpto. Ciencias de la Computación e Inteligencia Artificial*
*Universidad de Sevilla*
Sevilla, España
Correos electrónicos UVUS y de contacto (si distinto)

***Resumen*—El objetivo principal de este trabajo es comparar tres algoritmos de aprendizaje por refuerzo: Q-Learning, Montecarlo y SARSA, en el contexto de la planificación de rutas para un robot móvil en un entorno con obstáculos. El estudio se centra en evaluar la eficiencia computacional, la convergencia y la robustez de cada algoritmo al navegar hacia un destino minimizando la posibilidad de colisión.**

**Los resultados obtenidos muestran que cada algoritmo presenta ventajas y desventajas específicas en diferentes aspectos del problema planteado. Q-Learning demostró una rápida convergencia en la mayoría de los escenarios, mientras que Montecarlo ofreció una mejor exploración del espacio de estados. SARSA, por su parte, destacó en entornos altamente estocásticos debido a su enfoque de aprendizaje on-policy. Estas conclusiones ofrecen una guía para la selección del algoritmo más adecuado según las características del entorno y los requisitos del sistema.**

***Palabras clave*—Inteligencia Artificial, Aprendizaje por Refuerzo, Q-Learning, Montecarlo, SARSA, Procesos de Decisión de Markov, Entornos Estocásticos, Política, Exploración y Explotación.**

## I. Introducción

El aprendizaje por refuerzo es un tipo de aprendizaje automático que se enfoca en la toma de decisiones en entornos dinámicos y no deterministas. En este tipo de aprendizaje, el agente interactúa con el entorno y recibe recompensas o penalizaciones en función de sus acciones. El objetivo es maximizar las recompensas y minimizar las penalizaciones para encontrar la política óptima.

Esta técnica tiene una gran variedad de aplicaciones entre las que se encuentran predicciones financieras, robótica, videojuegos, medicina o cualquier problema de optimización [2] [3].

En el contexto de nuestro trabajo, nos centraremos en la aplicación del aprendizaje por refuerzo a la planificación de rutas para robots móviles. En este problema, un robot con ruedas debe encontrar una ruta segura y eficiente en un entorno con obstáculos. Aunque este problema puede parecer simple a primera vista, la presencia de obstáculos, la estocasticidad en el efecto de las acciones y la necesidad de optimizar la ruta para minimizar el tiempo y los recursos hacen que sea un desafío significativo.

Para evaluar la eficacia de los algoritmos de aprendizaje por refuerzo en este contexto, diseñamos tres mapas diferentes que varían en tamaño y porcentaje de obstáculos. Estos mapas nos permitirán comparar y analizar el rendimiento de los algoritmos en una variedad de escenarios. Comenzamos nuestro trabajo con el mapa de ejemplo proporcionado en la propuesta del trabajo, que tiene dimensiones de 15x51.

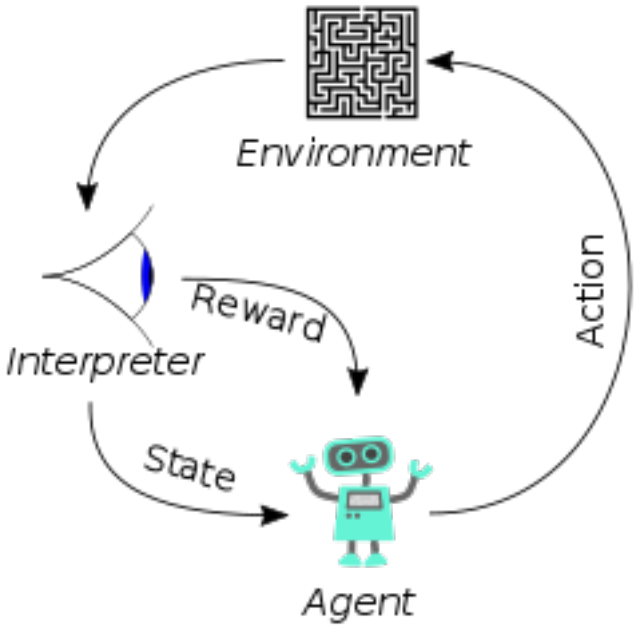


Fig. 1. Esquema del funcionamiento del aprendizaje por refuerzo.

## II. Preliminares

En esta sección, proporcionamos una introducción a los conceptos fundamentales del aprendizaje por refuerzo y la planificación de rutas de robots móviles en entornos con obstáculos, así como una descripción de los algoritmos de aprendizaje por refuerzo que utilizaremos en nuestro trabajo.

### A. Conceptos clave del aprendizaje por refuerzo

El aprendizaje por refuerzo (RL) es un tipo de aprendizaje automático que entrena al software para tomar decisiones y maximizar una recompensa en un entorno dado. Los componentes principales de un sistema de aprendizaje por refuerzo son:

- Agente: Es la parte del sistema que toma decisiones e interactúa con el entorno. En nuestro caso, el agente es el robot móvil que busca planificar rutas en un entorno con obstáculos.
- Entorno: Es el espacio en el que el agente se mueve y interactúa. En nuestro caso, el entorno es un mapa con obstáculos.
- Acción: Una acción es la decisión tomada por el agente en un estado particular. En nuestro caso, tenemos tenemos ocho acciones que representan cada uno de los posibles movimientos que puede efectuar el robot y una acción que representa que el agente permanezca en el mismo estado tras realizarla.
- Recompensa: La recompensa es la retroalimentación que el agente recibe del entorno después de tomar una acción en un estado dado. La recompensa puede ser positiva, neutra o negativa (esto sería una penalización).
- Política: La política es la estrategia utilizada por el agente para seleccionar acciones en función de los estados del entorno.
- Exploración: Es el proceso de explorar el entorno y aprender sobre las recompensas y penalizaciones.
- Explotación: Es el proceso de utilizar la información aprendida para tomar decisiones óptimas.

### B. Descripción de los algoritmos

A continuación, describimos los algoritmos de aprendizaje por refuerzo que utilizados en nuestro trabajo:

## III. Metodología

Esta sección se dedica a la descripción del método implementado en el trabajo. Esta parte es la correspondiente a lo realmente desarrollado en el trabajo, y se puede emplear pseudocódigo (nunca código), esquemas, tablas, etc.

A continuación, un ejemplo de uso de listas numeradas:

1) *Trabajos con dos alumnos:* poner nombre y apellidos completos de cada uno, y correos electrónicos de contacto (a ser posible de la Universidad de Sevilla). El orden de los alumnos se fijará por orden alfabético según los apellidos.
2) *Trabajo con un autor:* cambiar la cabecera de la siguiente manera
   a) *Una sola columna:* solo se debe especificar un alumno.
   b) *Información a añadir:* la misma que la especificada en el punto 1.

Las figuras se deben mencionar en el texto, como la Fig. **??**. También se pueden añadir ecuaciones, como la ecuación (1).

*mergesort*($V$)
**Entrada**: un vector $V$
**Salida**: un vector con los elementos de $V$ en orden
**si** $V$ es unitario **entonces**
    **devolver** $V$
**si no entonces**
    $V_1 \leftarrow$ primera mitad de $V$
    $V_2 \leftarrow$ segunda mitad de $V$
    $V_1 \leftarrow$ MERGESORT($V_1$)
    $V_2 \leftarrow$ MERGESORT($V_2$)
    **devolver** *mezcla*($V_1, V_2$)

*mezcla*($V_1, V_2$)
**Entrada**: dos vectores $V_1$ y $V_2$ ordenados
**Salida**: un vector con los elementos de $V_1$ y $V_2$ en orden
**si** $V_1$ no tiene elementos **entonces**
    **devolver** $V_2$
**si no si** $V_2$ no tiene elementos **entonces**
    **devolver** $V_1$
**si no entonces**
    $x_1 \leftarrow$ primer elemento de $V_1$
    $x_2 \leftarrow$ primer elemento de $V_2$
    **si** $x_1 \leq x_2$ **entonces**
        $x \leftarrow x_1$
        quitar el primer elemento de $V_1$
    **si no entonces**
        $x \leftarrow x_2$
        quitar el primer elemento de $V_2$
    $V \leftarrow$ *mezcla*($V_1, V_2$)
    añadir $x$ como primer elemento de $V$
    **devolver** $V$

Fig. 2. Algoritmo de ordenación `MergeSort`

$$a + b = \gamma \tag{1}$$

Un ejemplo de pseudocódigo se puede observar en la Fig. 2.

## IV. Resultados

En esta sección se detallarán tanto los experimentos realizados como los resultados conseguidos:

- Los experimentos realizados, indicando razonadamente la configuración empleada, qué se quiere determinar, y como se ha medido.
- Los resultados obtenidos en cada experimento, explicando en cada caso lo que se ha conseguido.
- Análisis de los resultados, haciendo comparativas y obteniendo conclusiones.

Se pueden hacer uso de tablas, como el ejemplo de la tabla I.

## V. Conclusiones

Finalmente, se dedica la última sección para indicar las conclusiones obtenidas del trabajo. Se puede dedicar un párrafo para realizar un resumen sucinto del trabajo, con los experimentos y resultados. Seguidamente, uno o dos párrafos

TABLA I
EJEMPLO DE TABLA

| A | B | C |
|---|---|---|
| 1 | 2 | 3 |
| 4 | 5 | 6 |

con conclusiones. Se suele dedicar un párrafo final con ideas de mejora y trabajo futuro.

## REFERENCIAS


[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] https://www.codificandobits.com/curso/aprendizaje-por-refuerzo-nivel-basico/2-ejemplos-reales-aplicacion-aprendizaje-por-refuerzo/

[3] https://www.aprendemachinelearning.com/aprendizaje-por-refuerzo/

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.