Explorar y describir cómo funciona el algoritmo de aprendizaje por refuerzo

Carlos Andres Suarez Torres

Saira Sharid Sanabria Muñoz

Universidad Santo Tomás

Facultad de Ingeniería Electrónica

Presentado a: DIEGO ALEJANDRO BARRAGAN VARGAS

Bogotá, Colombia

21 de noviembre de 2025

1. Exploración y Descripción del Funcionamiento del Algoritmo de Aprendizaje por Refuerzo

El algoritmo de aprendizaje por refuerzo funciona mediante un ciclo de interacción entre un agente y su entorno. En este proceso, el agente observa el estado actual del entorno y toma una acción basada en una política, que es una estrategia para seleccionar acciones. Como resultado de la acción tomada, el entorno responde proporcionando una nueva observación del estado y una recompensa, que puede ser positiva o negativa, según el resultado de la acción [1].

El aprendizaje por refuerzo puede dividirse en fases:

- Observación del estado actual del entorno.
- Selección y ejecución de una acción basada en la política.
- Recepción de la recompensa y nueva observación del estado.
- Actualización de la política para mejorar decisiones futuras.

Este ciclo de acción, observación, recompensa y ajuste se repite en múltiples episodios, permitiendo que el agente mejore continuamente su desempeño en el entorno [2].

1.1. ¿Cómo puede un agente aprender a tomar decisiones óptimas en un entorno incierto?

Un agente puede aprender a tomar decisiones óptimas en un entorno incierto mediante el aprendizaje por refuerzo, un método basado en la interacción directa con el entorno sin contar con un conocimiento previo del camino óptimo. El agente toma decisiones inicialmente al azar, luego observa las consecuencias de esas acciones en el entorno, y recibe retroalimentación en forma de recompensas o castigos. Con el tiempo, mediante ensayo y error, el agente ajusta su estrategia para maximizar la recompensa acumulada usando una política que define qué acción tomar en cada estado posible del entorno [3].

En entornos inciertos y dinámicos, donde no se conoce un modelo exacto de las consecuencias de las acciones, el agente puede funcionar sin un modelo interno del entorno, aprendiendo exclusi-

vamente mediante la experiencia directa con el ambiente. A medida que acumula experiencia, el agente mejora su capacidad para predecir qué acciones conducen a mejores resultados, incluso sin conocer todos los detalles del entorno. Este método es efectivo en problemas complejos como la navegación de robots, conducción autónoma o juegos donde el entorno cambia constantemente [4].

En resumen, el aprendizaje por refuerzo capacita a un agente para aprender decisiones óptimas en un entorno incierto a través de la experimentación interactiva, la evaluación de recompensas, y el ajuste continuo de su política de acciones para maximizar resultados a largo plazo.

1.2. Cuáles son los tipos de algoritmos de aprendizaje por refuerzo que existen y cuáles son sus arquitecturas ?, describir cada uno de los elementos

Existen varios tipos de algoritmos de aprendizaje por refuerzo (RL), que se pueden clasificar principalmente en dos grandes categorías: aprendizaje por refuerzo basado en modelos y aprendizaje por refuerzo sin modelos. Además, dentro de estas categorías hay distintos algoritmos que varían en cómo se representan y actualizan las políticas y funciones de valor [5].

1.2.1. Tipos principales de algoritmos de aprendizaje por refuerzo

1. **Aprendizaje por refuerzo basado en modelos:**
   - El agente intenta construir un modelo interno del entorno, aprendiendo las transiciones de estados y recompensas para simular resultados futuros.
   - Esto permite adoptar un enfoque planificado, anticipando consecuencias antes de actuar.
   - La arquitectura típica incluye un módulo para aprender o definir el modelo del entorno, otro para planificar la política de acción usando dicho modelo [6].
2. **Aprendizaje por refuerzo sin modelos:**
   - El agente aprende la política o función de valor directamente a partir de la interacción sin construir un modelo explícito del entorno.
   - Ejemplos incluyen Q-learning, SARSA y métodos de gradiente de política.

- La arquitectura suele incluir: un agente que evalúa estados y acciones mediante funciones de valor o políticas, y un mecanismo para actualizar estas funciones basadas en recompensas recibidas [5].

3. **Aprendizaje por refuerzo profundo:**

   - Usa redes neuronales profundas para aproximar las funciones de valor o políticas en ambientes con alta dimensionalidad y espacios de estados complejos.
   - Algoritmos populares son las Redes Q Profundas (DQN), PPO (Optimización de Políticas Proximales) y A3C (Asynchronous Advantage Actor-Critic).
   - Su arquitectura incluye redes neuronales que reciben como entrada el estado y generan acciones o valores Q, actualizadas mediante retropropagación y algoritmos de optimización.

1.2.2. Elementos comunes en la arquitectura de aprendizaje por refuerzo

- **Agente:** Entidad que toma decisiones y aprende para maximizar la recompensa.
- **Entorno:** El sistema con el cual el agente interactúa.
- **Estado (s):** Representación del entorno en un momento dado.
- **Acción (a):** Decisión o movimiento que el agente realiza.
- **Recompensa (r):** Señal de retroalimentación que indica el beneficio o costo de una acción.
- **Política ():** Función o regla que mapa estados a acciones.
- **Función de valor:** Estima la bondad de un estado o acción en términos de recompensa futura esperada.
- **Modelo del entorno (opcional):** Representación de cómo el entorno cambia en respuesta a las acciones (solo en los métodos basados en modelos).

Estos elementos trabajan en un ciclo continuo donde el agente observa el estado, selecciona una acción según su política, recibe una recompensa y un nuevo estado, y actualiza sus políticas o valores para mejorar su comportamiento a futuro **sutton2018reinforcement**.

1.2.3. Clasificación detallada de algoritmos

Los algoritmos de aprendizaje por refuerzo (RL) se pueden clasificar en varios tipos principales, cada uno con sus arquitecturas y elementos característicos:

1. **Algoritmos basados en valores**
   - Aprenden una función de valor que estima la utilidad de tomar una acción en un estado determinado para maximizar la recompensa futura.
   - **Q-learning:** Aprende la función Q, que estima el valor de la acción en un estado. Utiliza una tabla o función para almacenar estos valores y actualizarlos iterativamente con la fórmula de Bellman.
   - **SARSA (Estado-Acción-Recompensa-Estado-Acción):** Similar al Q-learning pero actualiza la función valor basada en la acción realmente tomada por la política actual, es decir, sigue la política durante el aprendizaje.
2. **Algoritmos basados en políticas**
   - En lugar de aprender una función de valor, estos algoritmos aprenden directamente una política que mapea estados a acciones, optimizando la toma de decisiones.
   - Ejemplos incluyen métodos de gradiente de políticas como REINFORCE y algoritmos más avanzados como PPO (Proximal Policy Optimization) y A2C (Advantage Actor-Critic).
3. **Algoritmos basados en modelos**
   - Crean un modelo del entorno, es decir, una aproximación de las transiciones de estado y las recompensas.
   - Con ese modelo, se pueden planificar estrategias antes de ejecutarlas en el entorno real, mejorando la eficiencia en el aprendizaje.
   - Permiten simular escenarios internos y evaluar políticas sin interactuar directamente [7].

4. **Aprendizaje por diferencia temporal (TD Learning)**
    - Combina ideas de los métodos basados en valores y basados en Montecarlo.
    - Aprende estimaciones basadas en aproximaciones sucesivas sin esperar al final del episodio.
5. **Reforzamiento profundo (Deep Reinforcement Learning)**
    - Usa redes neuronales profundas para aproximar funciones valor o políticas en problemas con espacios de estado de alta dimensionalidad.
    - Un ejemplo es el Deep Q-Network (DQN) que combina Q-learning con redes neuronales.

1.3. En la industria estos algoritmos para qué se utilizan?

En la industria, los algoritmos de aprendizaje por refuerzo se utilizan para optimizar decisiones y procesos en entornos complejos y dinámicos, donde es fundamental adaptarse a cambios y maximizar resultados a largo plazo.

1.3.1. Aplicaciones industriales principales

- **Optimización de procesos:** Se aplican para simular escenarios, evaluar estrategias y mejorar la gestión de inventarios, planificación de producción y rutas de transporte, lo que reduce costes y mejora eficiencia operativa [8].
- **Mantenimiento predictivo:** Los sistemas detectan patrones que indican posibles fallos o desviaciones en equipos o procesos, anticipando fallos para minimizar tiempos de inactividad y costos de mantenimiento, asegurando productos de alta calidad [9].
- **Finanzas:** Se usan para sistemas automatizados de trading e inversión que aprenden a tomar decisiones bajo condiciones cambiantes del mercado, optimizando beneficios y gestionando riesgos en tiempo real [10].

- **Ciudades inteligentes y logística:** Ajustan en tiempo real semáforos para reducir congestiones y emisiones, y optimizan rutas de distribución para transporte eficiente y bajo costo [11].
- **Robótica y movilidad:** Permiten que robots desarrollen habilidades motrices avanzadas, autonomía en terrenos complejos, y optimizan tareas industriales como clasificación, empaque y conducción autónoma (ejemplo: sistema Autopilot de Tesla).
- **Marketing personalizado y atención al cliente:** Mejoran motores de recomendación (Netflix, Yahoo), ajustan precios dinámicos (Amazon, Uber), y entrenan chatbots para atención más efectiva y segmentación precisa de clientes.

En resumen, el aprendizaje por refuerzo aporta valor en la industria al permitir que los sistemas tomen decisiones secuenciales bajo incertidumbre, optimizando procesos, ahorrando costos, aumentando la calidad y mejorando la experiencia del cliente a través de la adaptación continua basada en la interacción con el entorno.

REFERENCIAS BIBLIOGRÁFICAS

[1] A. Pérez, *¿Cómo funciona el aprendizaje por refuerzo?* [En línea]. Disponible: https://www.obsbusiness.school/blog/como-funciona-el-aprendizaje-por-refuerzo, [Consultado: 15-sept-2025], 2024.

[2] Fundación Bankinter, *Q-learning: el algoritmo del aprendizaje por refuerzo*, [Consultado: 15-sept-2025], nov. de 2023. dirección: https://www.fundacionbankinter.org/noticias/q-learning/.

[3] J. Murel, *¿Qué es el aprendizaje de refuerzo?* [Consultado: 15-sept-2025], IBM, 2023. dirección: https://www.ibm.com/es-es/think/topics/reinforcement-learning.

[4] Amazon Web Services, *¿Qué es el aprendizaje mediante refuerzo?* [Consultado: 15-sept-2025], 2023. dirección: https://aws.amazon.com/es/what-is/reinforcement-learning/.

[5] DataCamp, *Aprendizaje por refuerzo: una introducción con ejemplos en Python*, [Consultado: 15-sept-2025], jul. de 2024. dirección: https://www.datacamp.com/es/tutorial/reinforcement-learning-python-introduction.

[6] Interactive Chaos, *Algoritmos de aprendizaje por refuerzo*, [Consultado: 15-sept-2025]. dirección: https://interactivechaos.com/es/wiki/algoritmos-de-aprendizaje-por-refuerzo.

[7] Aprende Machine Learning, *Aprendizaje por refuerzo*, [En línea]. Disponible: https://www.aprendemachinelearning.com/aprendizaje-por-refuerzo/, [Consultado: 15-sept-2025], nov. de 2025.

[8] INESDI, *Aplicaciones empresariales del aprendizaje por refuerzo*, [Consultado: 15-sept-2025], sep. de 2024. dirección: https://www.inesdi.com/blog/aprendizaje-por-refuerzo/.

[9] J. Hernández, *Aprendizaje por refuerzo: aplicaciones prácticas en negocios B2B*, [Consultado: 15-sept-2025], ImpactoTIC, sep. de 2025. dirección: https://impactotic.co/inteligencia-artificial/aprendizaje-por-refuerzo-aplicaciones-practicas-en-negocios-b2b/.

[10] Emeritus, *Aprendizaje por refuerzo en la práctica: 10 ejemplos inspiradores para aprender*, [Consultado: 15-sept-2025], nov. de 2024. dirección: https://translate.google.com/translate?u=https%3A%2F%2Femeritus.org%2Fsg%2Fblog%2Fbest-reinforcement-learning-examples%2F&hl=es&sl=en&tl=es&client=srp.

[11] Fraunhofer IIS, *Aprendizaje por refuerzo para aplicaciones industriales*, [Consultado: 15-sept-2025], 2023. dirección: https://www-iis.fraunhofer-de.translate.goog/en/ff/lv/dataanalytics/auto.html?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc.