Optimización dinámica en prosumidores energéticos mediante PPO enmascarado

Víctor García Chico[1], Manuel Ramos Jiménez[1], Lucía Riera Presmanes[1], Roberto Salamanca Barrios[1], y Ana Belén Gil González[1]

Grupo de investigación BISITE, Universidad de Salamanca, Salamanca, España
{victorgarciachico, manuelramosji, luripres robertosala, abg}@usal.es

Resumen

Este trabajo implementa un algoritmo de aprendizaje por refuerzo profundo (DRL) con *Proximal Policy Optimization* (PPO) adaptado mediante enmascaramiento de acciones para gestionar sistemas prosumidores residenciales. El agente procesa 13 variables operativas (precios horarios bidireccionales, consumo, producción fotovoltaica, SoC batería, etc.) seleccionando entre 10 acciones discretas combinando carga/descarga y transacciones con la red, garantizando factibilidad física mediante máscaras dinámicas.

1. Introducción

Dado el alto consumo energético de la edificación (40 % en Europa), la gestión óptima de la generación fotovoltaica (DPV) y baterías (DBS) es crucial. Este problema de decisión secuencial es ideal para el Aprendizaje por Refuerzo (RL). Este trabajo presenta un agente de RL para un prosumidor basado en PPO [1], cuya contribución principal es una técnica de enmascaramiento de acciones. Este mecanismo restringe al agente a operaciones físicamente válidas, garantizando la seguridad y acelerando la convergencia del aprendizaje [2].

2. Metodología y resultados

Se desarrolló un entorno de simulación siguiendo el estándar de RL, que simula un ciclo anual completo mediante intervalos temporales de una hora (episodio de 8760 pasos). Para la producción solar se simularon los perfiles de producción de una instalación fotovoltaica de 48 m^2 en Salamanca utilizando el caso de uso PV-Inverter del software TRNSYS, empleando de datos para la simulación un año meteorológico típico obtenido de PVGIS. Respecto al consumo energético, se generaron perfiles de consumo anuales para un hogar de dos adultos trabajadores mediante la librería `pylpg`, una implementación de *LoadProfileGenerator*. Los precios de compra/venta de la energía se extrajeron del sistema de información eSIOS de Red Eléctrica de España. Para la batería se modeló una con 10 kWh de capacidad y una tasa de carga/descarga de 5 kW, similar a modelos comerciales. La batería modelada no presenta degradación por su uso.

El problema se formalizó como un Proceso de Decisión de Markov (MDP), con las siguientes características: en cada paso, el agente recibe un vector de observación con 13 variables que describen el estado del sistema: precios de compra/venta, consumo y producción actuales, demanda no cubierta, exceso de energía, nivel de la batería, costes e ingresos acumulados, e indicadores temporales (hora, día, mes). Se definió un espacio de 10 acciones discretas que combinan la gestión de la batería y la interacción con la red (e.g., .ªlmacenar energía sobrante y vender el exceso", "descargar batería y comprar para cubrir consumo"). Se aplicó un enmascaramiento de acciones para deshabilitar las acciones no viables según el estado actual, forzando al agente

a elegir solo entre opciones válidas. Para guiar el aprendizaje, la recompensa se define como el cambio en el beneficio neto entre el paso actual t y el anterior $t-1$.

El agente se entrenó utilizando el algoritmo *MaskablePPO* de la librería `sb3_contrib`. La política se modeló como una red neuronal de dos capas ocultas con 64 neuronas cada una. El entrenamiento se realizó con 3 perfiles de consumo y producción durante 10^7 pasos de tiempo (aprox. 1140 años simulados). Para la evaluación, el rendimiento del agente de RL se compara con un sistema *baseline* basado en reglas lógicas, que prioriza el autoconsumo (usa la batería para cubrir la demanda y almacena cualquier excedente) sin interactuar estratégicamente con la red.

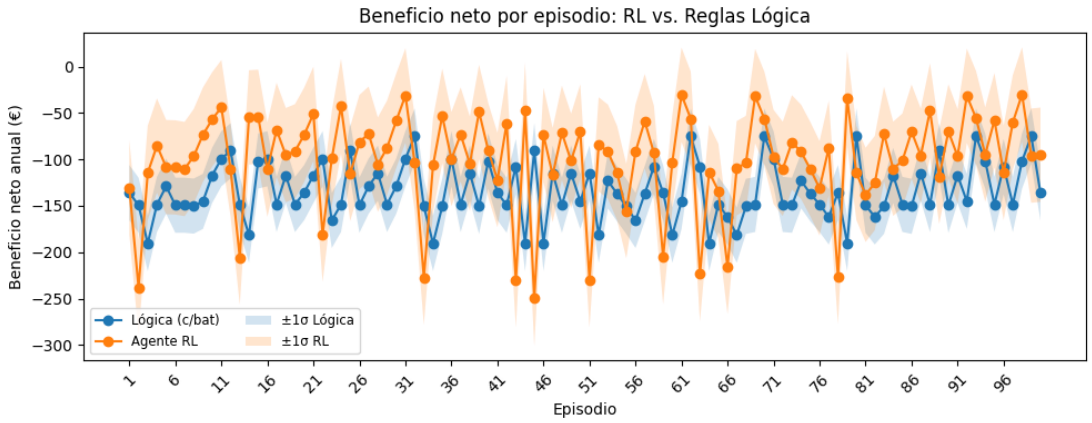

Figura 1: Evaluación del agente sobre 100 perfiles simulados de consumo (años).

En la evaluación, el agente de RL demostró una mejora significativa frente al *baseline*. En promedio, el coste neto anual se redujo de 134.85 € a 102.38 €, lo que supone un ahorro del 24 % (32.39 € anuales). El agente superó al sistema basado en reglas en el 79 % de los escenarios evaluados. Aunque en el 21 % de los casos tuvo un rendimiento inferior, la tendencia general muestra una clara ventaja económica.

3. Agradecimientos

Esta investigación cuenta con el apoyo de la subvención TSI-100933-2023-1 financiada por la Convocatoria de Cátedras Universidad-Empresa (Cátedras ENIA 2022) del Ministerio de Transformación Digital y Función Pública de España, y por el Plan de Recuperación y Resiliencia de la Unión Europea NextGenerationEU/PRTR.

Referencias

[1] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[2] Cheng-Yen Tang, Chien-Hung Liu, Woei-Kae Chen, and Shingchern D You. Implementing action mask in proximal policy optimization (ppo) algorithm. *ICT Express*, 6(3):200–203, 2020.