Enrique Ulises Báez Gómez Tagle
0241823@up.edu.mx Universidad Panamericana — Maestría en Ciencia de Datos - Aprendizaje de Máquina 2

18 de octubre de 2025

ML2: Flocking y Foraging Cooperativo mediante Aprendizaje por Refuerzo Multiagente

1. Descripción General del Problema

Este proyecto busca modelar el comportamiento colectivo de animales (aves o peces) que se mueven en grupo (flocking) y buscan alimento (foraging), simulando cómo los individuos mantienen cohesión grupal, evitan colisiones y consumen recursos de modo equilibrado en un entorno 2D.

El enfoque parte de reglas locales tipo Boids y extiende el modelo mediante aprendizaje por refuerzo (RL), permitiendo que los agentes optimicen sus decisiones y emerjan patrones adaptativos de cooperación y consumo sostenible.

2. Justificación y Objetivos

Justificación. El estudio de comportamientos emergentes en grupos animales permite entender principios aplicables a ingeniería, robótica de enjambre y sistemas distribuidos. Incorporar RL eleva el modelo tradicional de Boids hacia una representación más adaptativa y dinámica, acercándolo a los desafíos reales de inteligencia colectiva.

Objetivo general. Desarrollar una simulación de flocking y foraging cooperativo usando aprendizaje por refuerzo multiagente que equilibre cohesión grupal, búsqueda de recursos y distribución sostenible del consumo.

Objetivos específicos.

- O1: Implementar un entorno unificado de flocking y foraging con parches de recursos renovables.
- O2: Medir métricas de cohesión, alineación y eficiencia de consumo bajo reglas clásicas.
- O3: Introducir recompensas en RL que promuevan cooperación y eviten sobreexplotación.
- O4: Evaluar la distribución del consumo (índice de Gini) y la estabilidad del grupo con y sin RL.

3. Datos a Utilizar

El proyecto usará datos sintéticos generados por el entorno de simulación. Cada agente registra posición, velocidad, distancia a vecinos y nivel de recurso consumido. No se utilizan datos personales ni externos, y toda la simulación es controlada y reproducible. El dataset contendrá aproximadamente 1-2 millones de pasos de simulación, asegurando volumen suficiente para entrenar y validar los modelos de RL propuestos. El entorno de simulación será desarrollado completamente como parte de este proyecto, utilizando código propio para generar los datos sintéticos y controlar todas las dinámicas del sistema. No se emplearán entornos ni datasets externos, garantizando que el desarrollo, ejecución y recolección de datos sean 100 % originales y reproducibles.

4. Metodología y técnicas de Aprendizaje de Máquina

El modelo inicia con un sistema basado en reglas (Boids: cohesión, alineación y separación). Posteriormente se integra un agente de Aprendizaje por Refuerzo Profundo (PPO de Stable-Baselines3) con política compartida entre individuos. La recompensa combina términos de cohesión, éxito en la búsqueda de alimento y penalización por sobreexplotación o colisiones. Se probarán configuraciones de 6 a 20 agentes en entornos 2D con distintos niveles de densidad de recursos. Las

métricas de evaluación incluyen cohesión media, índice de alineación, Gini de consumo y tasa de regeneración de parches.

4.1. Definición del espacio de estado y de acciones

Espacio de estado. Cada agente observa un conjunto limitado de variables que describen su entorno local para mantener la dimensionalidad controlada. Las principales variables incluidas son:

- Posición relativa (x, y) y velocidad propia (vx, vy).
- Promedio de posiciones y velocidades de sus k vecinos más cercanos (cohesión y alineación locales).
- Distancia y dirección al parche de recurso más cercano.
- Nivel de recurso actual disponible en dicho parche (capacidad relativa 0–1).

En total, cada agente observa aproximadamente entre 10 y 14 variables continuas, con el objetivo de mantener el espacio de estado compacto y eficiente para el entrenamiento.

Espacio de acciones. Se emplea un conjunto discreto de acciones básicas que controlan el movimiento del agente:

- Girar a la izquierda o derecha ($\pm\Delta\theta$)
- Acelerar o desacelerar ($\pm\Delta v$)
- Mantener dirección actual

Cada agente tiene cinco posibles acciones discretas, con el objetivo de reducir el espacio de acción y evitar una explosión combinatoria, así es posible entrenar políticas efectivas con PPO sin comprometer estabilidad ni tiempo de cómputo.

4.2. Definición del ambiente

El ambiente es un espacio bidimensional continuo y limitado que representa el ecosistema donde los agentes (aves o peces) interactúan. Se disponen múltiples parches de recursos alimenticios, definidos por su capacidad máxima y tasa de regeneración. Los agentes perciben su entorno local, toman decisiones para moverse, alimentarse o mantenerse cohesionados, y el entorno aplica condiciones de frontera reflectantes. La dinámica se desarrolla en pasos discretos, generando observaciones, recompensas y un estado terminal al cumplirse los criterios de finalización.

5. Plan de Trabajo y Cronograma Preliminar

- **Semana 1:** Integración de código base (flocking + foraging) y visualización.
- **Semana 2:** Medición de métricas clásicas y penalización de sobreexplotación.
- **Semana 3:** Implementación de entorno Gymnasium y entrenamiento con PPO.
- **Semana 4:** Ajuste de recompensas y evaluación comparativa entre reglas y RL.
- **Semana 5:** Experimentos de cooperación y análisis de consumo distribuido.
- **Semana 6:** Presentación de resultados, y videos de comportamiento emergente.

6. Resultados y Posibles Retos

Se espera observar comportamientos emergentes de cooperación y autoorganización, donde los agentes mantengan cohesión mientras buscan y comparten recursos. Con RL, el grupo debería reducir colisiones y mejorar la distribución del consumo entre los parches disponibles. Los resultados incluirán videos de simulaciones, gráficas de métricas y comparativas entre modelos con y sin RL.

6.1. Riesgos y mitigaciones

- **Riesgos.** El principal riesgo es que el entrenamiento se vuelva lento o ineficiente al incrementar el número de agentes o la complejidad del entorno. También podría presentarse una falta de

convergencia si las recompensas no están bien equilibradas entre cooperación y consumo individual.
- **Mitigaciones.** Comenzar con un número reducido de agentes y complejidad baja en el entorno; ajustar gradualmente las recompensas para estabilizar el aprendizaje; emplear políticas compartidas para mejorar la eficiencia y monitorear las métricas de convergencia de forma continua.

6.2. Producto final esperado

- Dashboard o visualización de métricas de grupo (cohesión, alineación, consumo, Gini).
- Videos demostrativos de las simulaciones.

7. Bibliografía o referencias preliminares

- Reynolds, C. (1987). Flocks, herds, and schools: A distributed behavioral model. **ACM SIGGRAPH Computer Graphics**.
- Schulman, J. et al. (2017). Proximal Policy Optimization Algorithms.
- Stable-Baselines3 Documentation. https://stable-baselines3.readthedocs.io/
- Couzin, I. D. (2003). Self-organized animal groups and collective behavior. **Trends in Ecology & Evolution**.