Control de múltiples robots por aprendizaje operante: evitación de obstáculos, cooperación y formaciones espaciales.

D. A. Gutnisky & B. S. Zanutto

Instituto de Ingeniería Biomédica, FI-Universidad de Buenos Aires, Av. Paseo Colón 850, C1063ACV, Buenos Aires, Argentina. TE: (5411)4343-0891, Fax: (5411) 4786 2564, e-mail: dgutnisky@fi.uba.ar, silvano@fi.uba.ar.
Instituto de Biología y Medicina Experimental (IByME)– CONICET, Vuelta de Obligado 2490, C1428ADN, Bs. As., Argentina.

Resumen

La teoría de sistemas adaptativos, la cibernética y la psicología experimental fueron los pilares para incorporar al modelado de las funciones superiores del cerebro la dimensión temporal, los procesos a lazo cerrado y los mecanismos capaces de alcanzar objetivos predeterminados. En este abordaje al tema, sólo últimamente se han incluido algunos aspectos fundamentales del comportamiento animal. Dichos aspectos involucran el tiempo real y sistemas realimentados con la propiedad de alcanzar un objetivo por la interacción entre un sistema con capacidad de aprendizaje y el ambiente.
La biología evolutiva se interesó particularmente en la cooperación entre animales desde los tiempos de Darwin. Hamilton elaboró una teoría sobre la cooperación entre individuos relacionados genéticamente, y Trivers explicó cuales eran los mecanismos necesarios para la existencia de comportamientos cooperativos entre individuos no relacionados (altruismo recíproco). Trivers propuso que el juego del "Dilema del Prisionero" era un marco conceptual adecuado para comprender el altruismo recíproco. Si bien la cooperación entre animales es un área de estudio central en la biología evolutiva, ésta no se preocupa por los mecanismos biológicos involucrados en las tareas cooperativas. En el presente trabajo, mostramos que mediante mecanismos operantes se puede aprender a jugar el "Dilema del Prisionero" contra distintas estrategias aún en condiciones de ruido.
Por otro lado, recientemente ha habido un auge en las investigaciones en campo de la robótica sobre los llamados "sistemas multi-agentes". La cooperación en robots es de interés por varias razones, entre ellas, porque las tareas pueden ser demasiado complejas para que las lleve a cabo un sólo robot, o porque la eficiencia sería mayor si se utilizase varios de ellos de manera sinérgica. Sin embargo aún se carecen de teorías adecuadas para el desarrollo de dicha área.
El modelo propuesto asume que el animal puede predecir el estímulo incondicionado debido a que fueron identificadas neuronas dopaminérgicas del área tegmental ventral y en la sustancia nigra involucradas en el procesamiento de la predicción y el refuerzo. Las redes neuronales en que nos basamos tienen una neurona para cada posible respuesta de la red y otra para predecir el estímulo incondicionado. Las entradas de información de las neuronas que computan las respuestas son la memoria a corto plazo de los estímulos condicionados, incondicionados y la predicción. Los pesos sinápticos se calculan de acuerdo con la ley Hebbiana o anti-Hebbiana, dependiendo del valor de la predicción del estímulo incondicionado. Los pesos sinápticos de la neurona de predicción se calculan basados en la regla de Rescorla-Wagner. Finalmente, la red ejecutará cualquier respuesta mayor que cierto umbral.
En el presente trabajo se valida el modelo como alternativa en el control de robots en una tarea de evitación de obstáculos, comparando su performance con la del Q-Learning, uno de los algoritmos más utilizados en el ámbito de la Inteligencia Artificial. Para demostrar las capacidades del modelo para explicar cooperación entre agentes se utilizó como marco teórico el juego del Dilema del Prisionero, usualmente utilizado por los biólogos evolucionistas para explicar el altruismo reciproco. Finalmente, para analizar su aplicación en el ámbito de los sistemas multi-agentes, se realizaron simulaciones donde un grupo de robots debe aprender a realizar formaciones geométricas, con información local, sin necesidad de comunicación ni de líderes. Los resultados positivos encontrados tanto en su aspecto teórico como en su utilización en problemas de relevancia en Robótica e Inteligencia Artificial, sugieren que dichas áreas podrían beneficiarse del estudio del comportamiento animal, psicología experimental y neurociencias.

1. Introducción

El diseño de sistemas adaptativos es una de las claves en la investigación de la ingeniería en el presente siglo. Frank[7] afirma que los ingenieros enfrentan desafíos similares a los que se encuentran en los sistemas biológicos, como el problema de almacenamiento, complejidad e impredicibilidad. En los últimos años científicos e ingenieros han utilizado sistemas adaptativos para resolver problemas de la ingeniería.

Para solicitar copias del trabajo comunicarse con el Dr. B. Silvano Zanutto, Instituto de Ingeniería Biomédica, FI-Universidad de Buenos Aires, Av. Paseo Colón 850, C1063ACV, Buenos Aires, Argentina. TE: (5411)4343-0891, e-mail: silvano@fi.uba.ar.

Brooks[4] afirma que la IA clásica tiene sus bases en fundamentos defectuosos y que por eso existe una gran brecha entre el camino plausible de la inteligencia humana y el componente digital equivalente. En los últimos años ha habido un creciente interés de los investigadores en Inteligencia Artificial por lograr que los robots sean capaces de lograr cumplir una cierta tarea sin necesidad que la manera de resolverla sea programada previamente. En este sentido, el Aprendizaje Reforzado ha sido planteado como un marco teórico atractivo para ser utilizado para el control de agentes móviles. El aprendizaje reforzado, especialmente la programación atrajo a gran cantidad de investigadores, principalmente desde el trabajo de Watkins[18] donde desarrolló el Q-Learning. Touretzky y Saksida [16] dicen que los robots entrenados en las técnicas que utiliza Q-Learning no logran la sofisticación y versatilidad de los animales. Aunque parte de dicha diferencia en capacidades se debe a la superioridad perceptual y capacidades motoras de los animales. Ellos sugieren que se debería poner más atención en la literatura de aprendizaje animal y realizar intentos serios por modelar los efectos descriptos y que así los investigadores de la robótica podrían obtener interesantes beneficios. Por otro lado, los sistemas biológicos muestran un alto grado de flexibilidad para resolver una gran variedad de diferentes tareas. La investigación en el ámbito de la Robótica enfrenta problemas similares a los que tienen los animales en su lucha por su supervivencia. El condicionamiento operante, una forma de aprendizaje animal, es similar al aprendizaje reforzado en el sentido que permite a un agente adaptar su comportamiento para obtener premios del entorno (como ser comida, agua, etc) cuando realiza una acción correcta.

Los *animats*, robots autónomos o simulaciones de animales, están motivados por las limitaciones percibidas en la IA clásica. Dean[6] afirma que los animats son un intento para comprender la capacidad de los animales para generar de manera autónoma comportamientos adaptativos en entornos complejos y cambiantes. El comportamiento adaptativo se entiende de mejor manera focalizándose en la interacción entre el individuo y su entorno. Los conceptos para el desarrollo de animats, son tomados de la etología, sicología, neurobiología, de modelos formales y de la biología evolutiva.

Cao et al.[1] sostiene que ha habido un incremento en el interés en investigar sistemas de varios robots que muestren comportamientos cooperativos. La cooperación en robots es de interés por varias razones, entre ellas porque las tareas pueden ser demasiado complejas para que las lleve a cabo solo un robot, o porque la eficiencia sería mayor si se utilizase varios robots de manera sinérgica. Otra razón es que hacer varios robots más sencillos puede ser más barato, fácil, flexible y tolerante a fallas que construir un único robot poderoso para cada tarea.

La perspectiva con la cual se encara la problemática de realizar robots con capacidades y flexibilidades mayores a las propuestas por las líneas tradicionales del control automático nos ha llevado a primero demostrar la validez de esta nueva forma de encarar la problemática, para luego ser capaces de ir brindando soluciones cada vez más complejas y cercanas a la aplicación en maquinas inteligentes en campo. Primero se demuestra que el modelo de comportamiento operante desarrollado en nuestro laboratorio es capaz de hacer las tareas típicas que realizan los algoritmos de la robótica tradicional. Se tomó el problema de lograr que un móvil aprenda a esquivar obstáculos mediante aprendizaje reforzado, por ser una de las tareas más usuales en la robótica. Comparamos el modelo de condicionamiento operante ya desarrollado en nuestro laboratorio Zanutto y Lew[19] con uno de los algoritmos más utilizados en el control de robots, Q-Learning, y demostramos que nuestro modelo obtiene una mejor performance.

Se estudió el problema de la cooperación desde la perspectiva de la biología evolutiva y la teoría de juegos, para lo cual debimos cuestionar los modelos actuales por haber relegado el papel del aprendizaje en la cooperación entre individuos no relacionados, ya que si bien la cooperación entre animales es un área de estudio central en la biología evolutiva, ésta no se preocupa por los mecanismos de aprendizaje involucrados en las tareas cooperativas. Uno de los factores esenciales para comenzar a proponer modelos de cooperación es intentar responder la pregunta de cuál es la mínima capacidad cognitiva necesaria para que un agente aprenda a cooperar con otros. Se propone que el condicionamiento operante es una condición suficiente para aprendar a cooperar en el Dilema del Prisionero Iterado, utilizando como teoría de aprendizaje operante al modelo neuronal presentado por Zanutto y Lew[19]. Estudios previos realizados en nuestro laboratorio demuestran el papel fundamental del aprendizaje operante en la cooperación, abriendo así un nuevo camino para su aplicación a robots que aprenden a cooperar.

Al ser el modelo capaz de aprender a cooperar en el mencionado paradigma, nos preguntamos si sería capaz de resolver tareas de cooperación en un sistema multiagente. La tarea elegida fue la de formación espacial, donde un grupo de robots debe ser capaz de conformar una distribución geométrica específica y mantenerla, por ser una de los problemas típicos planteados por los investigadores de Inteligencia Artificial. De esta manera, se demuestra que el modelo de condicionamiento operante es capaz de lograr de manera emergente formar y mantener formaciones espaciales con otros robots.

En este trabajo está organizado de la siguiente manera. En la sección II se explica los fundamentos básicos del aprendizaje operante. En la sección III se presenta el modelo de condicionamiento operante.

En la sección IV se explica brevemente el algoritmo del Q-Learning. En la sección V se presentan los resultados de la comparación entre el modelo y el Q-Learning en una tarea de evitación de obstáculos. En la sección VI se analizan los resultados obtenidos en el juego del Dilema del Prisionero. En la sección VII se muestran los resultados de la tarea de formación espacial en un sistema multiagentes. Finalmente la sección VIII está dedicada a discutir los resultados generales del trabajo, y en la sección IX se presentan brevemente las conclusiones.

2. Aprendizaje Operante

Los psicólogos han propuesto al condicionamiento clásico y al operante como dos formas fundamentales de aprendizaje, que les permite a los animales adquirir características relevantes de su ambiente para conseguir refuerzos o evitar castigos. El condicionamiento operante es un procedimiento experimental a circuito cerrado, en el sentido que los estímulos recibidos por el animal son contingentes en su comportamiento. El animal aprende a realizar las acciones que lo conducen a abstener recompensas y evitar las acciones que le traen como consecuencia un castigo. Por ejemplo, una rata puede ser entrenada para presionar una determinada palanca cuando ve una luz roja (estímulo condicionado, CS), que le permite recibir una recompensa en forma de alimento (estímulo incondicionado, US).

3. Modelo de Condicionamiento Operante

Zanutto y Lew[19] presentaron un modelo basado en redes neuronales del condicionamiento operante para estímulos apetitivos y aversivos. Desde bases neurobiológicas y de comportamiento, se asume que los animales tienen la capacidad para computar una predicción del estímulo incondicionado (US). Dicha predicción controla el aprendizaje de las neuronas de respuesta para obtener refuerzo en el caso apetitivo y evitarlo en el caso aversivo.
La siguiente figura es un esquema de la red neuronal propuesta:

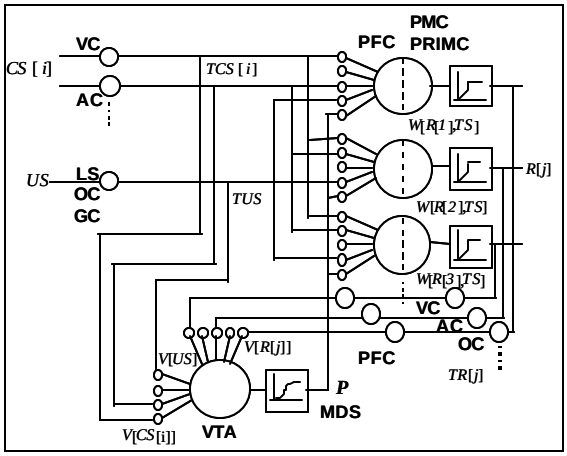

Figura 1. *Esquema del modelo Zanutto-Lew (MZL).*

Las entradas al modelo son todos los estímulos condicionados (CS), y el estímulo incondicionado (US). Existe una salida para cada respuesta posible. Los distintos bloques funcionales que componen la red son: 1) Trazas de estímulos y respuestas. 2) Neuronas de salida. 3) Neurona de predicción.
El objeto del modelo propuesto por Zanutto y Lew es encontrar hipótesis simples y biológicamente plausibles que puedan explicar el comportamiento apetitivo y aversivo basándose en las observaciones experimentales en animales.
Los experimentos de comportamiento sugieren que el aprendizaje está controlado por la expectación de los futuros eventos. Tanto en condicionamiento clásico como operante, el estímulo incondicionado (CS), anticipa al estímulo incondicionado (US). Rescorla y Wagner[12], propusieron que los animales aprenden comparando lo que esperan en una dada situación y lo que reciben. Schultz et al.[13] encontraron que existen sustratos neuronales involucrados en la predicción y el refuerzo tales como los que están involucrados las neuronas dopaminérgicas del área vetra tegmental (VTA) y la de la sustancia nigra.
El funcionamiento y las hipótesis involucradas en la arquitectura del modelo son sencillos. La neurona de predicción recibe como entradas las trazas de todos los estímulos y de las respuestas efectuadas. Los pesos sinápticos de todas las entradas salvo del US son plásticos (modificables), el peso de US se lo fija a un cierto valor, que desde el punto de vista de la teoría de los dos factores produce miedo para el caso

aversivo y sensación de saciedad en el caso apetitivo. La neurona de predicción tiende a predecir el US, mediante la utilización de la regla delta (formalmente igual a la de Rescorla-Wagner). La salida de la neurona de predicción está conectada a las neuronas de respuesta, que además tienen acceso a la información de las trazas de los CS y del US. Si la salida de la neurona de predicción supera un cierto umbral, el aprendizaje en las neuronas de salida se hará mediante la regla hebbiana en el caso apetitivo y antihebbiana en el aversivo, y de manera inversa cuando la salida de la neurona de predicción se encuentre por debajo del umbral. Cuando alguna de las neuronas de salida supera un umbral, se ejecuta la respuesta asociada.

3.1 Trazas

Existen dos tipos de trazas, las correspondientes a los estímulos, y las que corresponden a las respuestas. Tanto en el caso de las respuestas como el de estímulos condicionados, las trazas tienen la función de representar una memoria de corto plazo.
Las trazas de los estímulos reciben como entrada a los mismos estímulos, que pueden provenir de la corteza visual (VC), auditiva (AC), olfativa (OC), gustativa (GC) o del sistema límbico (LS). La salida de las trazas de corto plazo, ingresa tanto a las neuronas de salida, como a la neurona de predicción. Las trazas de las respuestas reciben como entrada a la salida de las neuronas de respuesta, y entran como realimentación a la neurona de predicción. La ecuación que genera las trazas de corto tiempo (T_S) de los estímulos (S), tanto condicionados (CS), como incondicionados (US) es para el instante n (Ec: 6.1-6.2):

$$T_{S_n} = T_{S_{n-1}} \cdot (1-\varepsilon) + \varepsilon \cdot S_{n-1} \text{ si } S_{n-1} > 0$$

$$T_{S_n} = T_{S_{n-1}} (1-\beta) \qquad \text{si } S_{n-1} = 0$$

La ecuación que determina la traza de las respuestas es:

$$T_{R_n} = T_{R_{n-1}} \cdot (1-\beta) + \varepsilon \cdot (1 - T_{R_{n-1}}) \cdot R_{n-1} \tag{6.3}$$

3.2 Neurona de Predicción

Las entradas a la neurona de predicción son todas las trazas de tiempo corto de los CS, la de US, y la de todas las respuestas (R). La salida de la neurona de predicción (P) entra a cada una de las neuronas de respuesta, y además tiene la función de controlar el aprendizaje de las neuronas de respuesta.

$$X_n = V_{US_n} \cdot T_{US_n} + \sum_{i=1}^{N_{CS}} V_{CSi_n} \cdot T_{CSi_n} + \sum_{i=1}^{N_R} V_{Ri_n} \cdot T_{Ri_n} \tag{6.7}$$

$$P_n = \frac{\chi}{1 + e^{-u \cdot (X_n - s)}} \tag{6.8}$$

Siendo P la función de salida, V los pesos y T las trazas correspondientes al US, los CS, y las respuestas (R). La cantidad de estímulos condicionados lo denotamos por N_{CS} y la cantidad de respuestas por N_R.
El peso V_{US_n} se mantiene fijo en un valor preestablecido. La actualización de los pesos de la neurona de predicción se basa en el modelo de Rescorla-Wagner (formalmente equivalente a la Regla Delta).

$$VX_{S_n} = VX_{S_{n-1}} + \eta(US) \cdot T_{S_n} \cdot (US_n - X_n) \tag{6.9}$$

$$V_{S_n} = \frac{2}{1 + e^{-k \cdot VX_{S_n}}} - 1 \tag{6.10}$$

con $\eta(US) = \eta_i$ cuando US>0 y $\eta(US) = \eta_d$ cuando US=0. Se utiliza la ecuación (6.10) para mantener los pesos limitados entre –1 y 1. A la vez a los pesos VX no se los deja crecer por encima de 10 ni decrecer por debajo de –10.

3.3 Neuronas de Salidas

Existe una neurona de salida por cada respuesta posible. Las salidas quedan determinadas por:

$$R_n(j) = g(Y_n(j))$$

$$Y_n(j) = W_{j\,\mathrm{Pred}_n} \cdot P_n + W_{jUS_n} \cdot T_{US_n} + \sum_{i=1}^{N_{CS}} W_{jCSi_n} \cdot T_{csi_n} + ruido(n) \tag{6.13}$$

$g = 0$ si $Y_n(j) < 0; g = 1$ si $Y_n(j) > \grave{\imath}$; sino $g = Y_n(j)$

El animal ejecuta una respuesta R(j) si Y(j) es mayor que el umbral m. En la implementación computacional, sólo se permite la ejecución de una respuesta por vez. La actualización de la salida de las neuronas no se hace de manera sincrónica sino que se elige al azar una sola neurona para que actualice su salida.

La ecuación que determina el aprendizaje de las neuronas de salida, está basado en la regla de Hebb. Si la neurona de predicción predice un US (supera cierto umbral) en el caso apetitivo el aprendizaje será Hebbiano (la asociación de estímulos y respuesta provocó que se obtenga un premio, y por lo tanto dicha asociación debe fortalecerse) y anti-Hebbiano en el caso que sea aversivo (la asociación de estímulos y respuesta provocó el castigo, y por lo tanto se la debe disminuir). Los pesos se actualizan sólo en la neurona que ejecutó la respuesta.

$$W_{jq_n} = y \cdot W_{jq_{n-1}} + f \cdot \Omega \cdot T_{Q_{n-1}} \cdot T_{R_{n-1}}(j) \qquad (6.17)$$

Siendo Q la entrada correspondiente (predicción, CS, o US). En el caso apetitivo si $P < l$ entonces $\Omega = -l$ y si $P \geq l$ entonces $\Omega = l$, de manera inversa en el caso aversivo.

5. Evitación de obstáculos

El objetivo de esta sección es comparar un modelo obtenido a partir de investigaciones en disciplinas tales como biología, sicología y neurociencias, con algoritmos utilizados en el aprendizaje de robots. El objetivo es acercar un modelo cuyo comportamiento básico es el que realizan muchos animales, al campo de la robótica; mostrando que puede resolver problemas para los cuales la mayoría de los investigadores en el área de la robótica prefieren las técnicas de la programación dinámica.

Se realizó una aplicación para controlar el desplazamiento de un móvil con sensores de cercanía en un entorno con obstáculos. El móvil contiene 5 sensores digitales, y tres posibles acciones, avanzar y girar a izquierda o derecha 45 grados y la capacidad de detectar si chocó contra un obstáculo. Las acciones se efectúan en tiempos discretos, una vez ejecutada una acción el móvil se desplazará una distancia fija. Los giros son mediante avance y rotación (existe una traslación del centro de masa), teniendo igual velocidad de giro que de avance, el ángulo final será 45º mayor o menor dependiendo para donde se haya realizado el giro. El control del móvil lo realiza el modelo que se elija, las posibilidades son el MZL, o el Q-Learning.

El móvil tiene un largo y un ancho de 10 cm. Existen 3 sensores en el frente del vehículo, situados equidistantemente y mirando al frente, y dos sensores en los costados de adelante, orientados lateralmente (ver Figura 3). Para poder calcular con mayor precisión cuando se ha producido un choque, la distancia fija a recorrer se divide en pasos más pequeños, pero no se devuelve información de los sensores en dichos instantes. La figura 4 muestra el recorrido que efectúa el móvil para cada una de las posibles acciones, y calculando 5 pasos intermedios. Cuando se avanza derecho la distancia recorrida es de 5cm, mientras que cuando se gira, se lo hace con un radio de 5cm, haciendo que el arco recorrido por el centro del vehículo sea de 3.93cm de longitud. Los sensores tienen un determinado ángulo de apertura (60º), y la energía que reciben es proporcional a la energía que irradiaría los obstáculos capturados por la visión del sensor, considerando a los obstáculos de manera bidimensional. Los sensores son digitales, y dan como salida un "1" lógico cuando el nivel percibido por el sensor supera un cierto umbral. El valor elegido de umbral hace que el sensor de un "1" lógico cuando tiene un obstáculo a 5cm del frente, y ocupa toda su visión.

Planteamos dos maneras distintas de procesar la información de los sensores. La manera más directa para utilizar el agente de Q-Learning y para el control del móvil, es codificar la información digital de los 5 sensores, en 32 estados. La otra manera que hemos elegido (que para el MZL es la única que se emplea), es realizar una lógica de los 5 sensores, devolviendo sólo 3 bit de información. Cada uno se activará cuando se tenga una visión más libre para la derecha, izquierda o para adelante, con la restricción que sólo uno de los bits permanecerá activo a la vez, por lo tanto frente a la situación de tener visión libre para dos de los tres lados, se optará por sólo una dirección, y en el caso de no tener ningún obstáculo obstruyendo la visión, los 3 bits permanecerán apagados (ver figura 4).

Las simulaciones consisten de 150 experimentos donde en cada uno de ellos se vuelve a comenzar desde la posición inicial. El experimento termina cuando se completan 400 movimientos o se ha chocado contra un obstáculo. Para realizar un análisis estadístico de la performance de los distintos agentes, se realizan 50 corridas de las características descriptas.

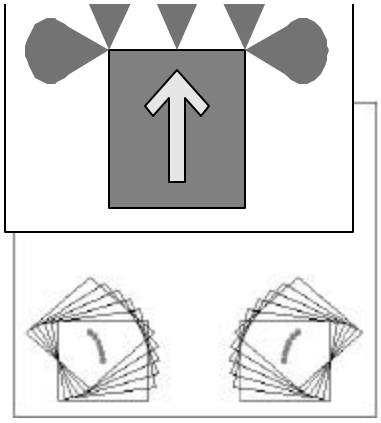

Fig. 2. Movimientos permitidos.

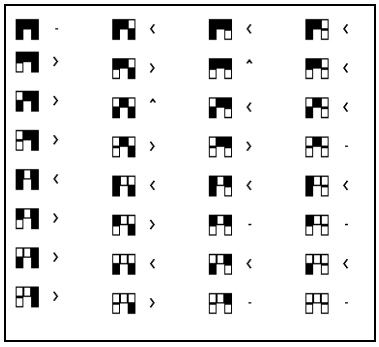

Fig. 3. Distribución de los sensores.

Fig. 4. Lógica de los 5 sensores y su correspondiente señal de salida.

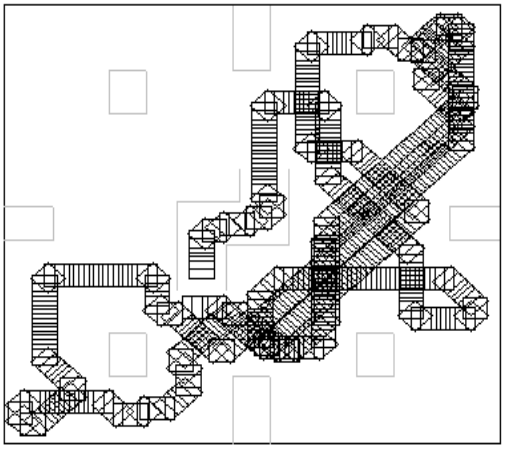

Fig. 5. Recorrido realizado por el modelo ZL una vez que aprendió a evitar obstáculos

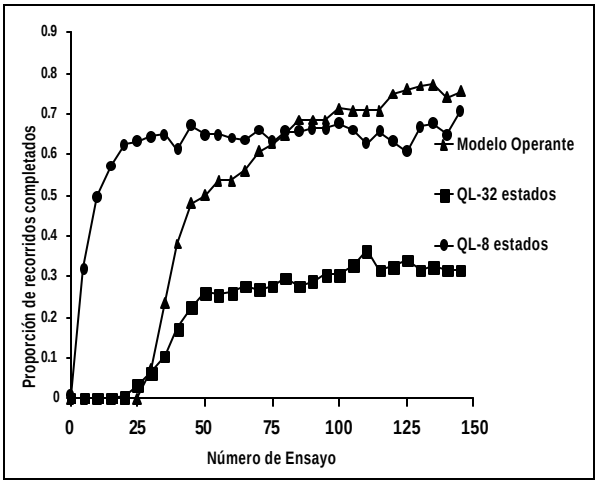

Fig. 6. Performance comparada de los tres tipos de agentes utilizados.

6. Cooperación y el Dilema del prisionero

Una de las posibles formas de cooperación en la naturaleza es el altruismo recíproco, intercambiar actos altruistas donde el beneficio es mayor que el costo, produce que a través del tiempo ambos individuos obtengan un beneficio neto. El problema radica en que un individuo siempre está tentado a no cooperar, para obtener más ganancia. Si los individuos interactúan de manera infrecuente se supone que prevalecerá la estrategia egoísta, pero en caso que existan muchos posibles encuentros, aquel que no coopere sufrirá ciertas pérdidas si los altruistas responden a los egoístas dejando de cooperar con éstos. Por eso son seleccionados aquellos que se protegen de sí mismos de los individuos no reciprocantes. Trivers[17], con su influyente trabajo sobre altruismo recíproco, fue quien relacionó este tipo de cooperación en la naturaleza con la teoría de juegos, más específicamente el dilema del prisionero.

El dilema del prisionero fue creado por Flood y Dresher, pero fue popularizado en 1953 por Von Neumann y Morgenstein. Dentro del campo de las matemáticas ha sido estudiado dentro de la llamada "teoría de juegos", y es uno de los juegos más estudiados dentro de la microeconomía.

En el juego participan dos jugadores que pueden optar entre dos posibles decisiones: cooperar o desertar. Ambos jugadores emiten su decisión simultáneamente, y en base a la acción conjunta de ambos, cada jugador obtiene un determiando puntaje (ver Tabla 1).

Se dice que dos jugadores adoptan un equilibrio de Nash si cada uno emplea una estrategia que es la mejor respuesta contra la del oponente. De esta manera si el dilema del prisionero se juega una sola vez, el único equilibrio de Nash es la mutua defección, porque se obtiene más puntaje más alla de lo que el oponente elija. Como resultado, ambos jugadores obtienen 1 punto en vez de 1, si hubiesen cooperado mutuamente.

	C	D
C	3	0
D	5	1

Tabla 1.

Genéricamente podemos establecer la relación que deben tener los distintos pagos, para que el juego mantenga las características enunciadas:

1) DC > CC > DD > CD

2) $2 \cdot CC > DC + CD$

Esta última condición determina que la decisión óptima en conjunto es cooperar, en este caso en particular se sumarían 6 puntos, mientras que si uno coopera y el otro no, se sumarían 5 puntos, y ninguno de los dos cooperando 2 puntos. El juego no tiene otra solución que la paradoja planteada, sin embargo, permitiendo que los jugadores se encuentren en reiteradas oportunidades (juego del dilema del prisionero iterado) surgen interesantes análisis de la conveniencia o no de ciertas estrategias de juego.

En el juego del dilema del prisionero iterado, los jugadores se enfrentan en sucesivas rondas. Para lograr maximizar el puntaje obtenido, los jugadores pueden cambiar sus jugadas dependiendo de la estrategia del oponente. Si la cantidad de rondas no se sabe de antemano, no existe ninguna estrategia que sea la mejor sin importar el comportamiento de las otras estrategias.

A partir del trabajo de Axelrod y Hamilton[1] sobre la cooperación entre individuos no relacionados, el dilema del prisionero volvió a cobrar vigor. Utilizando simulaciones por computadora, evaluaron la performance de un conjunto de estrategias en el caso del dilema del prisionero iterado, buscando estrategias evolucionalmente estables. Lo que difirió en el trabajo de ellos con respecto a los anteriores, son que el modelo presentado era nuevo en el tratamiento probabilístico de la posibilidad de encuentros entre dos individuos. Consideraron además no sólo la estabilidad final de una dada estrategia, sino la viabilidad de éstas en un entorno dominado por estrategias no cooperativas, y la robustez que presentaban en una variedad de entornos compuestos por distintas proporciones de estrategias.

Encontraron que si la probabilidad de encuentro de un dado agente estaba por encima de un umbral, además del éxito de la estrategia de siempre desertar (ALLD), aparecía otra estrategia robusta, llamada “Tit for Tat” (TFT) que en un movimiento inicial coopera, y luego realiza la misma acción que hizo su oponente en la jugada anterior. En el torneo organizado por Axelrod, el juego consistió de catorce estrategias, y donde los enfrentamientos entre cada una de ellas eran de 200 movimientos. El ganador fue la estrategia TFT. Este torneo luego fue modificado, permitiendo además una simulación de evolución[2], y donde fueron incluidas 64 estrategias. Luego de una primera ronda inicial, la segunda generación se compone en proporción al éxito en la ronda inicial. Nuevamente el ganador fue TFT.

A partir de los trabajos de Axelrod, apareció una gran cantidad de investigaciones en este campo, introduciendo interesantes variantes. Una de las variantes más interesantes fue la posibilidad de permitir a los jugadores “equivocar” su estrategia, ya que es más natural pensar que los animales cometen ciertos “errores”, haciendo que las conclusiones sobre las estrategias cambien notoriamente. Por ejemplo dos jugadores enfrentándose con la misma estrategia TFT, presentan una gran vulnerabilidad, ya que si uno equivoca un movimiento y pasa a no cooperar, el otro jugador no cooperará en la jugada siguiente, produciendo una alternancia entre cooperar y desertar, disminuyendo así de manera significante la performance de TFT.

En el presente experimento, se evalúa el comportamiento del modelo de condicionamiento operante, en el juego del Dilema del Prisionero Iterado. Para ello se lo enfrenta contra estrategias usuales dentro del juego, ya utilizadas en los enfrentamientos realizados por Axelrod[1][2]. Cada par de agentes se enfrenta en 1000 oportunidades, y para obtener estimaciones de los valores medios obtenidos, se realizaron a la vez 100 repeticiones en las mismas condiciones iniciales. Se realizan a la vez enfrentamientos con cierta probabilidad de hacer alguna respuesta al azar, más allá de la estrategia definida por el agente en cuestión. Esta posibilidad de “equivocarse”, fue analizada por Boyd[3]. Se mide así la robustez de cada una de las tácticas, y logrando que el análisis en la perspectiva del altruismo recíproco sea más realista, ya que difícilmente se encuentren casos donde los animales tengan una conducta sin ningún componente aleatorio.

Se utilizaron algunas de las estrategias más conocidas del dilema del prisionero. A continuación se describen brevemente cada una de ellas.

Tit For Tat (TFT) : En el primer movimiento coopera, y luego repite la jugada anterior del adversario.

All Coperate(ALLC) : Coopera siempre.

All Defect (ALLD,): Deserta siempre.

Alternate Defect and Cooperate (ALTDC): Alterna entre desertar y cooperar.

Modelo Zanutto-Lew (ZL) : Su respuesta depende del comportamiento de la red neuronal.

Azar : Elige las respuestas al azar, con probabilidad 0.5 para cada posibilidad.

Soft Majority (S-MAJO) : En el primer movimiento coopera, luego coopera, si el adversario ha cooperado igual o mayor cantidad de veces de las que ha desertado.
Slow TFT (S-TFT) : En los primeros dos movimientos coopera, luego si él cooperó en el movimiento anterior, cooperará a menos que el oponente haya desertado dos veces consecutivas, si en cambio en la jugada anterior había desertado, luego desertará a menos que el adversario haya cooperado dos veces seguidas.
Pavlov : En el primer movimiento coopera, luego coopera si en el movimiento anterior ambos agentes han hecho la misma acción.
Tit for two Tat (TF2T) : En las primeras dos jugadas coopera, luego coopera salvo que el oponente haya desertado dos veces seguidas.
Llamamos estrategias sencillas a aquellas que efectúan su estrategia sin fijarse lo que realiza el oponente, y por el contrario llamamos estrategias complejas, a aquellas que condicionan su accionar, a la historia de jugadas del oponente (y posiblemente propias). Las estrategias sencillas son : ALLD, ALLC, ALTDC, y AZAR. Mientras que las complejas son: TFT, MZL, S_MAJO, S_TFT, PAVLOV y TF2T.
Los resultados de los puntajes obtenidos por el modelo de condicionamiento operante (ZL) contra cada estrategia para las distintas probabilidades de jugadas al azar se muestran en las siguientes figuras.

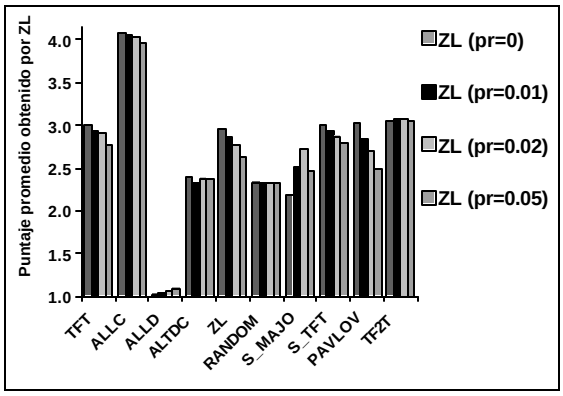

Fig. 7. Puntaje obtenido por el modelo ZL contra cada una de las estrategias para diferentes pr.

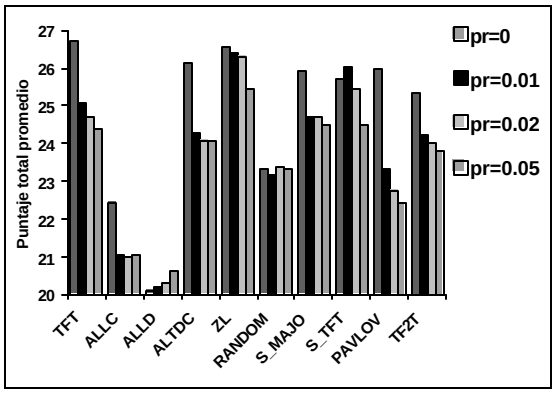

Fig. 8. Puntaje total obtenido por cada una de las estrategias para diferentes pr.

7. Formaciones espaciales en un sistema multiagente

Se estudió el problema de lograr un comportamiento global en un grupo de agentes simulados, únicamente mediante información local y una señal de reforzamiento. El objetivo global es establecer y mantener determinadas formaciones geométricas (una fila o una columna). Cada robot recibe un refuerzo dependiendo únicamente de su acción individual y no de una situación global específica. De esta manera, el modelo de condicionamiento operante logra cumplir la tarea, aprendiendo por ensayo y error y sin la necesidad de tener un control centralizado o comunicación explícita con los otros agentes. Al igual que en la propuesta de Mataric[11], no existen líderes y los agentes no son diseñados especialmente para que cooperen. Más aún, la cooperación emerge a raíz del diseño de la tarea y de las señales de reforzamiento.
Se desarrolló una aplicación donde un grupo de robots homogéneos se podían mover libremente, recibiendo recompensas dependiendo de condiciones locales. El entorno es una cuadrícula de 5 X 5, donde ningún agente puede ocupar la misma celda. Cada uno de los cinco robots recibe 5 señales a manera de entrada: Norte, Sur, Este, Oeste y "Doble unión". Las primeras cuatro indican la dirección donde hay mayor cantidad de robots. La señal "Doble unión" se recibe cada vez que el agente está alineado con otros dos agentes, uno a cada lado del mismo, o cuando está al lado de una pared y de otro agente (ver figura 9). Los robots tienen 5 posibles acciones, moverse en cada una de las 4 posibles direcciones, o reafirmar la "Doble unión" (mantenerse en la misma posición). En cada paso los robots deciden que acción tomar dependiendo de qué movimientos tenga permitido (no puede moverse a una celda ya ocupada, y sólo puede ejecutar la acción "Doble unión" si ésta señal está presente). Los agentes siempre tienen que moverse, no pueden quedarse quietos, a menos que estén completamente bloqueados.
Los agentes se mueven por turnos elegidos al azar. Si luego de que un agente ejecuta un movimiento y queda adyacente a otro robot, el primero recibe un refuerzo (un US en el modelo). Si el agente no se mueve y queda adyacente a otro robot, no recibe ningún refuerzo. La otra posibilidad para obtener refuerzos es recibir la señal de "Doble unión" y tomar la acción de reafirmarla. La única posible manera de lograr una formación estable es que todos los agentes tomen la acción de "Doble unión" cada vez que tengan robots adyacentes en ambas direcciones.

Se compararon la performance del modelo operante, con agentes que realizan movimientos al azar (salvo la acción de "Doble unión" que la efectúan siempre que sea posible) y con una lógica programada que siempre se mueve en la dirección donde hay mayor cantidad de robots. Los resultados muestran que los agentes controlados por el modelo operante aprenden la tarea de manera tal de obtener más refuerzos (figura 10) y que una vez aprendida la tarea, su performance alcanza a la de la lógica programada.

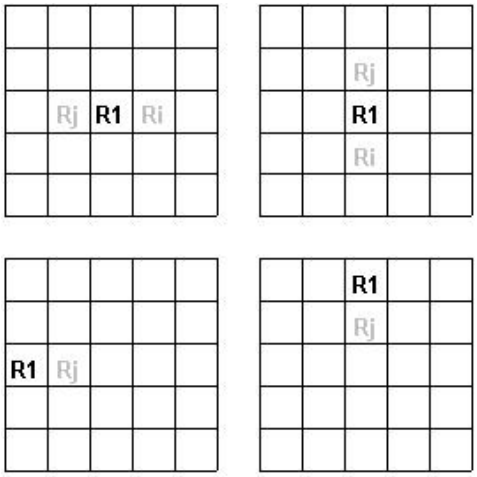

Fig 9. Condiciones en las cuales R1 obtiene una señal de "Doble Unión"

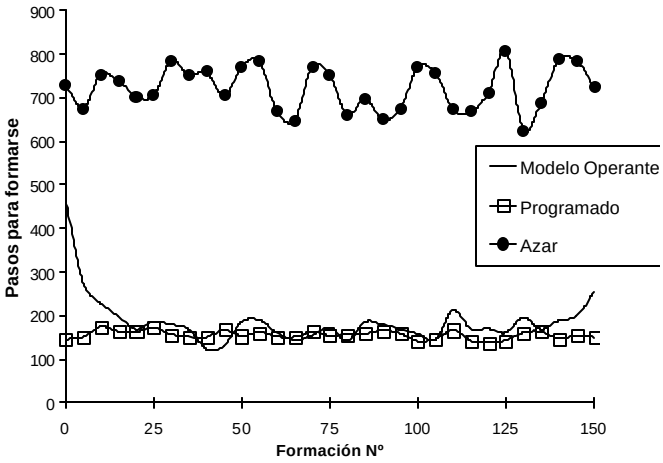

Fig. 9. Cantidad media de pasos necesarios para lograr una formación en función del número de formaciones ya realizadas.

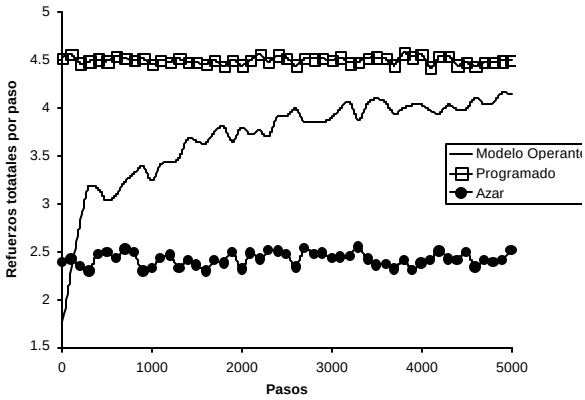

Fig. 10. Cantidad de refuerzo recibido por cada tipo de agente en función de la cantidad de pasos realizados.

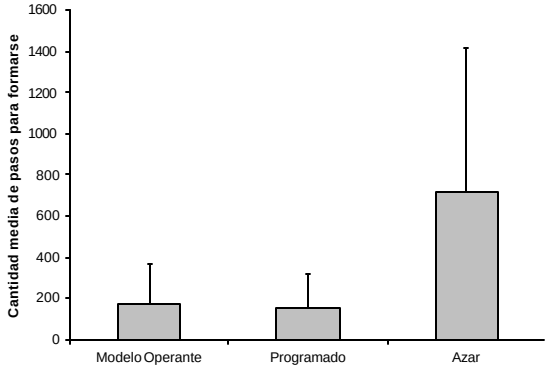

Fig. 11. Cantidad media de pasos necesarios para lograr una formación para cada tipo de agente una vez aprendida la tarea.

8. Discusión

Nuestra intención en la presente investigación está en la línea descripta por Touretzky y Saksida[16] y por las investigaciones en animats, el objetivo ha sido acercarse a la robótica y a su problemática, desde modelos sustentados en investigaciones de sicología, biología y de las neurociencias. Esta trabajo se ha basado en un modelo neuronal del comportamiento aversivo y apetitivo cuyas hipótesis de construcción fueron tomadas de las teorías de comportamiento, resultados experimentales en animales y evidencias neurobiológicas.

Sharkey[14][15] dice que la nueva corriente en robótica está formada por la coalición de distintos tipos de investigaciones y motivaciones, por un lado algunos lo están para probar un modelo neuronal, biológico o sicológico o incluso para probar una teoría general de comportamiento adaptativo, por otro lado están aquellos que quieren desarrollar robots sencillos y eficientes y utilizan el trabajo realizado en los sistemas naturales como una inspiración no restringida.

La definición de qué se entiende por cooperación, depende mucho del campo que la analice. En robótica existen muchas posibles definiciones, sin embargo nosotros tomamos la óptica de la biología evolutiva, más específicamente lo que se llama altruismo recíproco. En lo que respecta a la cooperación en general

los aspectos teóricos están recién en sus comienzos. La perspectiva que brinda la cooperación en animales no relacionados provee de una serie de fundamentos para utilizar y constatar la teoría que la sostiene y así estar en concordancia con la filosofía en este trabajo, que es sugerir analizar los modelos teniendo en cuenta los fundamentos teóricos tomados de distintas disciplinas y que ya han sido investigados con rigor científico.

9. Conclusiones

En este trabajo se ha mostrado que un modelo del comportamiento operante inspirado por datos neurobiológicos es capaz de realizar tareas habituales en el ámbito de la robótica. Se mostró que el modelo operante tiene una mejor performance en una tarea de evitación de obstáculos que el Q-Learning. Se demostró que mediante mecanismos operantes se podía lograr a aprender a cooperar, según el esquema del dilema del prisionero. Finalmente, en una tarea donde un grupo de robots recibían recompensas por mantenerse unidos, agentes controlados por nuestro modelo operante, eran capaces de lograr de manera emergente mantener determinadas formaciones espaciales. Estos resultados sugieren la utilización de modelos inspirados en comportamiento animal, para la solución de problemas en el ámbito de la robótica, de manera genérica, sin necesidad de adaptar cada algoritmo para cada problema en particular.

10. Referencias

[1] Axelrod, R. & Hamilton, W.D. (1981). The evolution of cooperation. *Science* **211**: 1390-1396.

[2] Axelrod, R. (1984). *The Evolution of Cooperation*. New York, NY: Basic Books.

[3] Boyd, R. (1989). Mistakes allow evolutionary stability in the Repeated Prisoner's Dilemma game. *Journal of Theoretical Biology, **136,*** 47-56.

[4] Brooks, R. A (1990). Elephants don't play chess. *Robotic and Autonomous Systems,* **6**:3-15.

[5] Cao, Y., Fukunaga, A. S., Kahng, A. B. & Meng, F. (1995). Cooperative Mobile Robotics: Antecedents and Directions. *In `IEEE/TSJ International Conference on Intelligent Robots and Systems', Yokohama, Japan*.

[6] Dean, J. (1998). Animats and what they can tell us. *Trends in Cognitive Sciences,* **2:2** , 60-67.

[7] Frank, S. A. (1996). The design of natural and artificial adaptive systems. In M.R. Rose & G.V. Laude (Eds.) *Adaptation*, 451-505. New York, N.Y.: Academic Press.

[8] Hamilton, W.D. (1964). The genetical evolution of social behavior I. *J. Theor. Biol.* **7**: 1-16.

[9] Hebb, D. O. (1949). *The organization of Behavior: A Neuropsychological Theory*. New York: Wiley.

[10] Lew, S. E., Wedemeyer, C., & Zanutto, B. S. (2001). Role of unconditioned stimulus prediction in the operant learning: a Neural network model. *Proceeding of IEEE* (Conf. on Neural Networks), pp 331-336.

[11] Mataric (1992). Designing emergent behaviors: From local interactions to collective intelligence. *From Animals to Animats 2, Second International Conference on Simulation of Adaptive Behavior (SAB92)*, , pp. 432-441.

[12] Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory*. New York: Appleton-Century-Crofts.

[13] Schultz, W., Dayan P. & Montague, R. (1997). A neural substrate of prediction and reward. *Science* **275**: 1593-1598.

[14] Sharkey, N. (1997). The new wave in robot learning. Robotics and Autonomous Systems, **22**(3-4).

[15] Sharkey, N. E. and Ziemke, T. (1997). A consideration of the biological and psychological foundations of autonomous robotics. In Sharkey, N. E. (Ed.), Robotics and Autonomous Systems, 22(3-4). *Special issue on Robot Learning: The New Wave*, Connection Science, **10**(3-4).

[16] Touretzky, D. S. & Saksida, M. L. (1997). *Operant conditioning in skinnerbots*. Adaptive Behavior, **5**(3/4):219--247.

[17] Trivers, R.L. (1971). The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**: 35-57.

[18] Watkins, C. J. C. H. (1989) Learning with Delayed Rewards. Ph.D. dissertation, Cambridge University, Psychology Department.

[19] Zanutto, B. S. & Lew, S. (2000). A neural network model of aversive behavior. In M.H. Hamza (Ed.) *Proceedings of the IASTED Neural Networks NN'2000*, pp. 118-123. IASTED/ACTA Press. Anaheim, Calgary, Zürich.