Pavlovian-instrumental interaction in 'observing behavior'

Ulrik R. Beierholm, Peter Dayan

2010

Las "irracionalidades" abundan en la conducta animal. Estos fallos en la optimalidad han sido fuente de refinamiento y revisión teórica.

Una tarea en que se observa tal desviación es un tipo de conducta de observación: los sujetos pueden recibir una recompensa grande o pequeña, cuyo tamaño se determina de forma estocástica. Cuando deben elegir entre averiguar antes o después cuál de las dos recompensas recibirán, los sujetos prefieren enterarse antes, a pesar de que el conocimiento no afecte al resultado final y a pesar de que esa elección sea costosa. En economía esta anomalía se expresa como la "resolución temporal de la incertidumbre" y se explica con constructos como el *savoring*.

Algunas explicaciones, como el "deseo de ganar información Shannon" han sido descartadas dado que los animales prefieren observar *más* a pesar de que el número de bits recibidos al hacerlo sea *menor*.

En un estudio reciente se registró la actividad de supuestas neuronas dopaminérgicas en el cerebro de monos mientras estos decidían observar. Una teoría común indica que estas neuronas reportan un error de predicción de diferencias temporales de las recompensas futuras, como en explicaciones de aprendizaje por refuerzo de la elección instrumental óptima. Se encontró que la actividad de las neuronas dopaminérgicas estaba asociada con la elección hecha por los monos a pesar de que, dado que la conducta de observación no ofrece ningún beneficio instrumental, no puede tener ningún error de predicción. Esto podría indicar que las neuronas dopaminérgicas podrían estar reportando aspectos del beneficio de la búsqueda de información además del reforzamiento primario.

Se examinará el grado en que esta forma de conducta de observación puede ser explicada por aprendizaje de diferencias temporales sumado a influencia pavloviana sobre acciones instrumentales. Se asume que los sujetos solamente hacen predicciones asociativas cuando se encuentran adecuadamente involucrados (*engaged*) en la tarea. Si el nivel de involucramiento es influido por el tamaño de las predicciones (el efecto supuestamente pavloviano) entonces los estímulos que predigan recompensas grandes certeras o deterministas llevarán a mayor involucramiento (atención). Se demuestra que los fallos ocasionales de involucramiento, modelados como una ruptura en el estado representacional de la memoria de trabajo, pueden llevar directamente a la preferencia por la observación y la

actividad dopaminérgica aparentemente anómala sin tener que incluir a la "información".

Método

Se modela el *value learning* mediante una versión modificada de un modelo estándar de diferencias temporales. Se asume que la tarea se puede especificar como un proceso de Markov, donde un participante estima la recompensa futura esperada (valor) para cada estado s como $V(s)$, actualizándolo de acuerdo con

$$V(s) \leftarrow V(s) + \alpha\delta V \tag{1}$$

donde α es la tasa de aprendizaje, y δV es el cambio en el valor esperado dado por:

$$\delta V = r + \gamma V(s) - V(s) \tag{2}$$

donde r es la recompensa entregada, y s' es el estado que sigue al estado s. El aprendizaje procede para los tres tipos de ensayo.

La única desviación del modelo estándar de diferencias temporales es que se asume que la actualización correcta del sistema depende del mantenimiento del involucramiento. Se asume que la probabilidad de desinvolucrarse en el curso del estado s es

$$\epsilon = \epsilon_0 \exp(-V(s)\psi) \tag{3}$$

por unidad de tiempo en segundos. Así, para un estado dado t la probabilidad de una actualización correcta es $1-P_{fail} = (1-\epsilon)^\tau$, donde τ es la cantidad de tiempo pasada en el estado. ϵ_0 y ψ son parámetros fijos. Se asume que la consecuencia del desinvolucramiento es la transición a un estado s^0 de valor $V(s^0) = 0$ (que no se actualiza), con lo que la actualización de la señal para $V(s)$ es

$$\delta V = r + \gamma V(s^0) - V(s) = -V(s). \tag{4}$$

El sistema permanece en ese estado hasta que se entrega una recompensa al final del ensayo. En ese punto el sistema se reinvolucra creando un error de diferencia temporal relativo al valor fijo de $V(s^0)$. Se asume que todo desinvolucramiento es negado al iniciar el siguiente ensayo.

La elección solo es posible en un estado C, entre pasar al estado C_D o a C_{ND}. Asumimos que el sujeto elige D basado en la función *softmax*

$$P(D) = \frac{exp(\beta V(C_D))}{exp(\beta V(C_{ND})) + exp(\beta V(C_D))} \tag{5}$$

Resultados

Se toman los datos de Bromberg-Martin y Hikosaka (2009).

Sujetos sedientos podían recibir con 50 % de probabilidad una cantidad pequeña o grande de agua. Había tres tipos de ensayos: *forced information* (se presentaba a los sujetos un solo objetivo y, al mirarlo, recibían una de dos claves que predecían el volumen de agua que recibirían), *forced-random* (se presentaba un solo objetivo y, al mirarlo, recibían uno de dos estímulos no correlacionados con el volumen de agua), y *free choice* (se presentaban ambos objetivos y los sujetos podían decidir entre ellos).

Los sujetos mostraron una preferencia gradual por la alternativa discriminativa a pesar de no haber diferencia en las ganancias esperadas entre alternativas.

Se construyó un modelo que involucra un algoritmo estándar de aprendizaje por diferencias temporales. Los ensayos forzados y libres permiten aprender sobre las ganancias esperadas, y después en los ensayos libres la elección depende de los valores relativos de cada alternativa mediados por una función *softmax*.

Cuando hay una demora entre la presentación de las claves y las consecuencias finales se debe asumir un mecanismo por el cuál los sujetos mantienen conocimiento de la tarea, es decir, memoria de trabajo. Existe el requisito mínimo de que los sujetos deben continuar involucrados en la tarea durante la demora para mantener esta memoria. Así, el modelo usado aquí se separa de modelos estándar de aprendizaje por diferencias temporales al asumir que el mantenimiento del involucramiento es influido por el valor predicho actual. Es decir, si el valor es alto, el involucramiento también.

Perder involucramiento afecta negativamente al desempeño de los sujetos, y por analogía con un efecto similar en el automantenimiento negativo este efecto es considerado pavloviano: una respuesta automática basada en predicciones apetitivas o aversivas.

En el modelo se considera que el involucramiento se pierde en algunos ensayos como una función estocástica del valor predicho que se actualiza. Esto tiene el propósito de disminuir el valor subjetivo de aquellos estados asociados con valores bajos más abajo aun de su valor objetivo. Esto afecta en mayor medida a las claves no discriminativas de la tarea. El desinvolucramiento asociado con un estímulo negativo es benigno dado que el resultado de estos ensayos se modela como cercano a cero. Esto en conjunto crea un sesgo a favor de elegir la alternativa discriminativa.

Bromberg-Martin y Hikosaka registraron la actividad de neuronas dopaminérgicas en el cerebro medio durante la tarea, y encontraron una activación consistente con la interpretación clásica de la función de estas neuronas: que reportan los errores de predicción de diferencias temporales en la entrega de recompensas futuras.

Sin embargo, hubo otra activación en el momento de aparición de los objetivos que indicaban un ensayo *forced-information* o *forced-random* que da pistas sobre la conducta de observación. El objetivo asociado con un ensayo *forced-information* tuvo un incremento leve en la actividad fásica, mientras que el objetivo que indicaba pistas aleatorias fue seguido por un leve decremento en la actividad. En la interpretación de diferencias temporales de estas neuronas, la activación es consistente con la preferencia de los monos, pero no con los valores objetivos.

Al aplicar el modelo de involucramiento variable en esta tarea se encuentra un patrón idéntico a los datos observados.

Al cambiar la consecuencia de apetitiva (agua) a aversiva (un choque eléctrico), cabía preguntarse cómo cambiaría el involucramiento: si sería dependiente de la saliencia o de la valencia (positiva o negativa). Los datos indicaron lo primero: una predicción de un castigo protegió al involucramiento.

Otra manipulación importante ha sido variar la probabilidad P_{rew} de la recompensa grande contra la pequeña. Mientras P_{rew} baja de 1 hacia 0.5, incrementa el sesgo a favor de la alternativa discriminativa. Más abajo de 0.5, el sesgo depende de la presunción sobre cómo se generan las elecciones. Una regla de elección que depende de la diferencia en los

valores esperados ($V_D - V_{ND}$) lleva a un sesgo que al final disminuye a 0 mientras lo valores mismos se acercan a 0. Si las elecciones se basan en la razón entre los valores ($\frac{V_D}{V_{ND}}$), el sesgo puede continuar incrementando mientras P_{rew} se acerca a 0. Tal incremento fue observado por Roper y Zentall mientras se adelgazaban los programas de reforzamiento.

Otro factor importante es que los sesgos inherentes al desinvolucramiento son pequeños y se desarrollan en tiempos largos. Esto significa que el curso de aprendizaje inicial puede ser sujeto a influencia de los valores iniciales adscritos a las opciones, llevando a sesgos que son desproporcionados con el estado final.

Discusión

Se ha mostrado una explicación de cómo la conducta de observación puede surgir de un pequeño sesgo pavloviano sobre la conducta instrumental asociado con el desinvolucramiento de la tarea sin recurrir a la "búsqueda de información". Se quiso capturar específicamente un experimento con macacos, pero los resultados tienen otros alcances.

Dinsmoor ha sugerido un fenómeno llamado "observación selectiva" en el que los sujetos se enfocan en los estímulos asociados con probabilidades más altas de recompensa. Esta explicación puede verse como una forma de observación selectiva, pero involucrando acciones internas asociadas con la asignación de involucramiento y atención, en lugar de acciones externas como la mirada preferencial.

A algunos les llama pensar que los animales buscan adquirir conocimiento sobre el mundo, pero las teorías de información tienen dificultades con los resultados de reducir la probabilidad de recompensa, lo que reduce la incertidumbre y la información ganada pero incrementa la observación.

En la explicación presente hay varias formas en que los valores predichos podrían influir en el involucramiento persistente. En ciertos experimentos con monos el entregar información que indica que la recompensa se encuentra decepcionantemente lejos hace que éstos se desinvolucren de la tarea. Lo mismo podría suceder aquí. El mecanismo más obvio para mediar esto sería la influencia de la dopamina sobre la memoria de trabajo. Otras teorías sugieren que la memoria de trabajo está controlada por un proceso de *gating* asociado con los ganglios basales.

Una pregunta interesante es en qué condiciones reemerge el involucramiento. Se pude asumir que sucede con la entrega de una recompensa, pero es necesario un mecanismo que lo formalice. Una manera de hacerlo sería mediante un re-involucramiento estocástico basado en el error de predicción de recompensa o en el valor esperado. Tal mecanismo podría suceder en cualquier momento pero sería mucho más probable en el momento de entrega de recompensa y en el inicio de un nuevo ensayo. También sería necesario especificar el caso de el desinvolucramiento en el momento de la selección de acciones: en un estado desinvolucrado el animal no ejecutaría ninguna elección, por lo que probablemente no respondería en el tiempo que tiene disponible. Si se requiere una elección para avanzar en el programa, esto tendría que suceder tras un re-involucramiento eventual.

En cuanto a reforzamiento condicionado, muchos autores consideran que los estímulos son reforzadores condicionados dada su asociación con la recompensa. En la perspectiva presente tanto S^+ como S_D y S_{ND} son reforzadores condicionados. La pregunta clave en conducta de observación es una concavidad aparente: el valor promedio de dos estímulos asociados de forma determinista con recompensas pequeñas y grandes es mayor que el valor

de un solo estímulo asociado de forma estocástica con las mismas consecuencias. Esta no-linealidad demanda explicación. Algunas perspectivas ponen el peso en el estímulo asociado de forma segura con la recompensa grande. La explicación presente pone el énfasis en los estímulos no discriminativos, sugiriendo que es más probable que lleven a desinvolucramiento.

Se ha mostrado que el efecto de observar, preferir estímulos discriminativos pero conductualmente irrelevantes, puede explicarse por una "mala conducta" pavloviana, lo que la pone en la misma categoría de otras suboptimalidades. Las explicaciones sobre teoría de la información son, aunque seductoras, innecesarias.