Information Seeking as Chasing Anticipated Prediction Errors

Jian-Qiao Zhu, Wendi Xiang, Elliot A. Ludvig

2016

Los organismos suelen preferir las opciones informativas, a menudo sacrificando las recompensas por resolver la incertidumbre antes.

Algunas variables que han mostrado ser críticas en la conducta subóptima son:

- La contingencia entre las claves predictivas y las consecuencias finales.
- La anticipación de la información con respecto a la consecuencia (demoras más largas llevan a mayor búsqueda de información).
- La magnitud de la recompensa potencial (mayor magnitud lleva a mayor valor subjetivo)
- Las consecuencias aversivas pueden producir evitación de la información (efecto *ostrich*).

El propósito de los autores es construir un modelo que capture tantos de estos resultados como sea posible.

Modelos computacionales existentes

Estas conductas desafían a los modelos estándar de *reinforcement learning* (RL). Se han propuesto extensiones a este marco de referencia, como el *information bonus model*, el *disengagement model*, y el *anticipatory utility model*.

El *information bonus model* propone que recibir información actúa como una recompensa primaria. El *anticipatory utility model* formaliza la idea de *savouring*, y propone que los animales disfrutan de la anticipación de una recompensa garantizada.

The Anticipated Prediction Error Model

Proponen un modelo centrado en la idea de los errores de predicción anticipada (*anticipated prediction errors*, APE). Según el modelo APE, los animales toman muestras *one-step* de los futuros anticipados de un modelo simple del mundo, y calculan el error de predicción asociado a la muestra. Estos errores de predicción son tratados como si fuesen una recompensa en sí mismos. Las muestras están sesgadas de modo que los futuros que contienen errores de predicción positivos tienen más probabilidad de ser muestreados.

Este *muestreo hacia adelante* (anticipación) desde el estado actual utilizando experiencias imaginadas y dinámicas ambientales aprendidas puede proveer señales anticipatorias útiles para la toma de decisiones.

La diferencia crítica entre el modelo APE y el modelo RL estándar es que APE tiene dos sistemas de valoración separados: uno estimado de la experiencia real (sin modelo) y otro estimado con este proceso de muestreo hacia adelante (basado en el modelo). Los errores de predicción son llamados *anticipated prediction errors*. Junto con la función convencional de valor, estos APEs guían la preferencia para buscar o evitar ciertos estados futuros. El sesgo hacia los errores positivos puede incluso producir elecciones subóptimas.

Especificación del modelo

Se extiende el modelo estándar de diferencia temporal (TD) en el que se asume que los agentes estiman una función de acción-valor para cada estímulo experimental:

$$Q(s_t, a_t) = \mathbb{E}[\sum_{k=1}^{\infty} \gamma^k r_{t+k-1}]$$

en donde t indica el tiempo, s_t indica el estado visitado en el tiempo t, r_t indica la recompensa inmediata entregada tras el tiempo t, y $\gamma \in [0, 1)$ es un factor de descuento, que devalúa las recompensas demoradas. Esta función representa la ganancia futura esperada descontada. Esta función se estima mediante un mecanismo simple de actualización:

$$Q\left(s_t, a_t\right) \leftarrow Q\left(s_t, a_t\right) + \alpha\delta_{t+1}$$

en donde α es la tasa de aprendizaje, y δ esel error de predicción de recompensa (RPE), calculado de la manera que sigue:

$$\delta_{t+1} = r_{t+1} + \gamma \max_a Q\left(s_{t+1}, a\right) - Q\left(s_t, a_t\right)$$

Esta señal de RPE representa la diferencia entre el valor del par actual estado-acción y el valor de el siguiente mejor par estado-acción más la recompensa lograda en la transición. Así, los RPE se disparan por cada transición de estado; los mecanismos del modelo APE dependen de la transición del estado de elección a las claves predictivas que revelan cuál de los resultados será entregado en un ensayo. Una buena clave que resuelva la incertidumbre resultará en RPEs positivos, mientras que una mala clave resultará en RPEs negativos. Los RPEs serán de cero como respuesta a claves no-predictivas, una vez que los valores estén bien aprendidos.

Se define a los *anticipated prediction errors* como la dicrepancia percibida entre el estado actual y un estado futuro anticipado. Formalmente, si no hay una recompensa primaria entregada durante la transición del estado s a s' (e.g., la transición del estado de elección a las claves) entonces el valor de APE en el estado s, cuando se anticipa el estado s', se define como el producto de los errores de predicción y la probabilidad de transición:

$$\mathrm{APE}\left(s' \mid s, a\right) = T\left(s' \mid s, a\right) \times \left[\gamma^{D_{ss'}} \max_{a'} Q\left(s', a'\right) - Q(s, a)\right]$$

donde $D_{ss'}$ es el tiempo que toma pasar de s a s', y $T(s' \mid s, a)$ es la probabilidad de transición de s a s' tomando la acción a.

Este cálculo depende de las muestras generadas basadas en la dinámica aprendida del ambiente. El principal supuesto del modelo AEP es que los organismos tratan a las APEs como recompensas: las APE positivas son reforzantes, y las negativas son aversivas. Las APE son positivas cuando el valor anticipado del futuro muestreado es mejor que el valor del estado actual, y son negativas en el caso contrario. Estas cantidades pueden entenderse como muestras del placer que uno deriva de anticipar la clave "buena", y el displacer de anticipar la clave "mala". Además, se le asigna un peso de atención a cada APE que especifica qué tan verosímil es que un futuro específico sea muestreado. Así, el valor de decisión $\bar{Q}$ de tomar la acción a se define como la suma ponderada de APEs para los resultados futuros anticipados más la función de valor para el estado correspondiente:

$$\bar{Q}(s,a) = \sum_{s_k \in \mathcal{S}} w_k APE\left(s_k \mid s, a\right) + Q(s,a)$$

donde $\mathcal{S}$ denota el conjunto de todos los posibles estados futuros tras tomar la acción a en el estado s a los que el sujeto puede atender.

Dados los valores de decisión de las alternativas con y sin clave, la función softmax (¿wtf?) se utiliza para computar la probabilidad de escoger la opción con claves:

$$P(a) = \frac{e^{\beta \bar{Q}(s,a)}}{\sum_{a' \in \mathcal{A}} e^{\beta \bar{Q}(s,a')}}$$

donde $\mathcal{A}$ es el conjunto de todas las acciones posibles en el estado s, y β es un parámetro de temperatura inverso, que controla el grado de exploración.

Experimento

Para probar su modelo, hicieron un experimento en el que humanos debían elegir entre una alternativa informativa y una no informativa con consecuencias positivas (imágenes eróticas), neutras (imágenes de objetos) o negativas (imágenes aversivas). Había ensayos buenos (consecuencias positivas y neutras), mixtos (positivas y negativas), malos (negativas y neutras). La predicción del modelo APE era que los participantes buscarían información solamente en los ensayos buenos y neutrales. La predicción viene del sesgo hacia el muestreo de posibles estados futuros positivos.

Resultados

Como se esperaba, los participantes escogieron la opción informativa por encima del azar en las condiciones buena y neutral, pero no en la mala, aunque hubo grandes diferencias individuales.

Comparación de los modelos

Se ajustó tanto el modelo APE como el *information bonus model* a las proporciones de elección.

Dado que una alternativa es no informativa, muestrear futuros esperados en ella no resulta en errores de predicción. Este análisis sugiere que en ese caso, solo los APEs relacionados con las claves predictivas determinan la elección en esta tarea particular.

Los modelos se ajustaron a los datos usando modelamiento jerárquico bayesiano. El modelo APE resultó ampliamente superior al modelo *information bonus*.

Discusión

El modelo APE asume que las preferencias son guiadas por errores de predicción anticipada (APEs) acumulados mediante la simulación de futuros posibles. Estos APEs son tratados como recompensas, lo cual, combinado con un sesgo para muestrear futuros con resultados favorables, lleva a la búsquea de información en situaciones con resultados positivos potenciales. El modelo predijo correctamente la tendencia de los sujetos a buscar información solamente cuando hay resultados favorables en juego.

El modelo APE además puede ser útil en otras situaciones de elección subóptima y búsqueda de información. Por ejemplo, el APE positivo escala con la probabilidad de recompensa (es más grande con probabilidades más bajas), lo que provee un mecanismo por el cual una alternativa con baja probabilidad de recompensa podría ser preferida sobre otra con mayor probabilidad. Además, el APE crece con la magnitud de las recompensas, lo que predice la mayor preferencia por opciones informativas observada en monos.

Este modelo se aplicó a una situación simple con muestreo de un solo paso. Situaciones más complejas con árboles de decisión más profundos serían computacionalmente imposibles. Pero de eso se encargarán otros.