MAT^2

MATerials MATemàtics
Volum 2006, treball no. 7, 8 pp. ISSN: 1887-1097
Publicació electrònica de divulgació del Departament de Matemàtiques
de la Universitat Autònoma de Barcelona
www.mat.uab.cat/matmat

Blade Runner, el "factor humano" y la fórmula de Bayes

Rosario Delgado de la Torre

El 27 de agosto de 2004 el periódico *The Guardian* publicó los resultados de una encuesta realizada a más de 60 científicos. En ella la película *"Blade Runner"*, de Ridley Scott (1982), fue escogida mayoritariamente como la mejor película de Ciencia Ficción de todos los tiempos (por delante de "2001, una odisea del espacio" de Stanley Kubrick, 1968), considerándola como realmente adelantada a su tiempo. Esta película está basada en la novela *¿Sueñan los androides con ovejas mecánicas?* del escritor norteamericano Philip K. Dick (1968).

La novela nos sitúa en un incierto futuro, concretamente en la sombría ciudad de Los Ángeles del año 1990. Rick Deckard es un "cazador de bonificaciones" de una agencia policial, que se dedica a capturar robots humanoides (androides) que llegan ilegalmente a la Tierra desde alguno de los planetas-colonia. Estos androides son fabricados por la Rosen Association, únicamente para ser utilizados como criados por los humanos que viven en los planetas-colonia. Algunos escapan a su destino y llegan ilegalmente a la Tierra (matando, en algunos casos, a los humanos que se interponen en su camino).

Las agencias policiales desean identificar a los androides que llegan ilegalmente a la Tierra para eliminarlos. ¿Cómo lo hacen? Se puede utilizar el análisis de médula, completamente fiable pero costoso, peligroso y lento. En su lugar, se usa una prueba (un test de empatía, cualidad que parece ser que

los humanos poseemos pero los androides no, por lo que resulta ser el "factor humano" a tener en cuenta). Si el resultado de aplicar el test a un individuo es *positivo* (+) éste se clasifica como *androide* (A), mientras que si es *negativo* (-) se clasifica como *humano* (H). Este test es barato, inocuo y rápido. Pero...no es completamente infalible. Puede fallar en dos sentidos: podría ser que un androide "pasase" el test, es decir, que el test de detección diese un *"falso negativo"*, de manera que no fuese identificado. Pero también podría resultar que un humano "no lo pasase" y fuese clasificado como androide (sería un *"falso positivo"*), con su consecuente eliminación. En palabras de Bryant, jefe de Rick, este hecho "sería lamentable".

Cuando la Rosen Association lanza al mercado una nueva serie de androides, los *Nexus-6*, con unidades cerebrales muy mejoradas, se plantea la cuestión de si para estos robots el último método de identificación desarrollado, el *test (modificado) de empatía de Voigt-Kampff*, es útil o no.

Se quiere decidir si se puede utilizar este test de manera generalizada, instalando controles policiales a tal efecto, de manera que los androides puedan ser identificados y eliminados. Naturalmente, para ello interesa conocer con qué asiduidad (es decir, con qué *probabilidad*) se producen los dos tipos de error en la identificación usando el test, el *falso positivo* y el *falso negativo*.

Por tanto, es necesario estimar las dos *características del test de Voigt-Kampff* que son las probabilidades de cometer los dos errores: la *probabilidad de falso positivo* y la *probabilidad de falso negativo*, digamos α y β, respectivamente. Para ello, se envía a Rick a probar el test en unos cuantos individuos a la Rosen Association. El resto de la novela no da información que permita hacer un estudio probabilístico del método de Voigt-Kampff de detección de androides, así que vamos a hacer un supuesto sobre los resultados que podría haber obtenido Rick, y veremos cómo se haría.

Imaginemos que Rick aplica el método de Voigt-Kampff a un total de 1000 individuos. Él no sabe si son androides o humanos, y todos ellos creen que son humanos, lo sean o no (a los androides se les ha instalado una unidad de memoria artificial a tal efecto), para evitar que el conocer la realidad influya en el resultado del test; es lo que se llama un *experimento de doble ciego*. Se anota el resultado del test para cada individuo: positivo o negativo. Posteriormente, la Rosen Association revela la verdadera identidad de los individuos del estudio, de manera que se puede saber cuántos individuos han sido correctamente clasificados. Recopilamos dicha información en una tabla de doble entrada como ésta:

	Humanos (H)	Androides (A)
+	7	282
-	693	18
Total	700	300

Y podemos calcular α =*probabilidad de falso positivo* = proporción de positivos de entre los humanos (lo escribimos $P(+/H)) = \frac{7}{700} = 0{,}01$ (en porcentaje, multiplicando por 100, es un 1 %). Es decir, aproximadamente la prueba da positivo (falsamente) en 1 de cada 100 humanos a los que se aplica el test. También podemos calcular β = *probabilidad de falso negativo* (lo escribimos $P(-/A))$ = proporción de negativos de entre los androides $= \frac{18}{300} = 0{,}06$ (en porcentaje, es un 6 %). La prueba da negativo falsamente en 6 de cada 100 androides a los que se aplica el test. Luego hemos estimado a partir de los 1000 individuos las características del test,

$$\alpha = P(+/H) = 0{,}01 \quad \text{y} \quad \beta = P(-/A) = 0{,}06$$

Teniendo en cuenta que $P(-/A) + P(+/A) = 1$ (aplicando el test a un androide, seguro que o bien da positivo o bien da negativo), y que de manera análoga $P(-/H)+P(+/H) = 1$, tenemos que $P(+/A) = 1-\beta$ y se denomina *"sensibilidad del test"* (es la probabilidad de clasificar correctamente a un androide). Análogamente, $P(-/H) = 1-\alpha$ y se denomina *"especificidad del test"* (es la probabilidad de clasificar correctamente a un humano). En este caso, la *sensibilidad* es de un 94 % y la *especificidad* de un 99 %.

Una vez hecho esto, Rick puede elevar el informe a sus superiores recogiendo estas características del test de Voigt-Kampff. ¿Qué decidirán ellos? ¿es o no adecuado el test para usarlo de manera generalizada en la población?

Antes de contestar a esta pregunta, deberíamos pensar qué queremos decir con "ser adecuado". Lo lógico es pensar que un test será adecuado cuando se da la siguiente circunstancia: si aplicamos el test a un individuo y da positivo (con la consecuente orden de eliminarlo), hay una alta probabilidad de que éste sea un androide (es decir, es poco probable que sea un humano).

Esta probabilidad, que es muy importante calcular, se llama *"valor predictivo positivo"*, y debería ser alta (estar próxima a 1 o, en porcentaje, al 100 %). Se da la "paradójica" circunstancia de que esta probabilidad NO es sólo una característica del test, como α y β, sino que depende del número o proporción de androides que haya en la población a la que se vaya a

aplicar. Esto es, el mismo test aplicado en diferentes poblaciones (con diferentes proporciones de androides en ellas) dará valores predictivos positivos diferentes.

Supongamos que el test se aplica en una ciudad donde la población humana es de 3000 individuos y se sabe que acaban de llegar de manera ilegal a ella 600 androides provenientes de Marte. Entonces, conociendo las probabilidades α y β asociadas al test, tenemos que aproximadamente el 1 % de los humanos darían positivo (esto es, unos 30), y el resto (2970) darían negativo, mientras que el 6 % de los androides darían negativo (unos 36), y el resto (564) darían positivo. Podemos resumir la información en esta tabla "ficticia" (lo es puesto que en realidad no aplicamos el test a toda la población de la ciudad; en cada casilla de la tabla se recoge el número de individuos que esperaríamos que hubiese en ella si lo hiciésemos):

	Humanos (H)	Androides (A)
+	30	564
-	2970	36
Total	3000	600

A partir de ella podemos calcular el valor predictivo positivo, que es la proporción de androides de entre los individuos que dan positivo,

$$\text{valor predictivo positivo } = P(A/+) = \frac{564}{564+30} = \frac{564}{594} \cong 0{,}94949495$$

(aproximadamente un 95 %). Entonces, la proporción de humanos de entre los individuos que dan positivo es $P(H/+) = 1 - P(A/+) \cong 0{,}05$. Esto nos dice que "únicamente" se eliminará por error (serán humanos) un 5 %, aproximadamente, de todos los individuos que den positivo en el test.

¿Qué cálculo hemos hecho, a partir de $\alpha = P(+/H)$ y $\beta = P(-/A)$, las características del test, y de la proporción de androides en la ciudad, $p = P(A) = \frac{600}{3000+600} = \frac{600}{3600} \cong 0{,}166667$, para obtener el valor predictivo positivo $P(A/+)$? Dividimos numerador y denominador por 3600, el número total de individuos en la ciudad, y el cálculo realizado se puede expresar de la siguiente manera:

$$P(A/+) = \frac{564}{564+30} = \frac{\frac{564}{3600}}{\frac{564}{3600}+\frac{30}{3600}} = \frac{\frac{564}{600}\,\frac{600}{3600}}{\frac{564}{600}\,\frac{600}{3600}+\frac{30}{3000}\,\frac{3000}{3600}} =$$

$$= \frac{P(+/A)\,P(A)}{P(+/A)\,P(A) + P(+/H)\,P(H)} \left(= \frac{(1-\beta)\,p}{(1-\beta)\,p + \alpha\,(1-p)} \right)$$

donde hemos usado que si $p = P(A)$ es la proporción de androides, entonces la proporción de humanos es $P(H) = 1 - p$. Esta fórmula se conoce como **Fórmula de Bayes**, en honor del matemático Thomas Bayes (1702-1761).

¿Qué habría pasado si el número de androides llegados a la ciudad hubiese sido mucho menor, 60 por ejemplo? El test sigue siendo el mismo, pero ¿lo será el valor predictivo positivo? La respuesta es que no. Ahora ya lo podemos calcular sin necesidad de construir una tabla "ficticia", usando la **fórmula de Bayes**, con α y β las mismas (pues son características del test), pero con el nuevo valor $p = P(A) = \dfrac{60}{3000+60} = \dfrac{60}{3060} \cong 0{,}0196078\,.$

Entonces, el valor predictivo positivo sería

$$\frac{(1-\beta)\,p}{(1-\beta)\,p + \alpha\,(1-p)} \cong \frac{(1-0{,}06)\,0{,}0196078}{(1-0{,}06)\,0{,}0196078 + 0{,}01\,(1-0{,}0196078)} \cong$$

$$\cong \frac{0{,}018431}{0{,}028235} \cong 0{,}652771$$

que es, en porcentaje, aproximadamente un 65 %. Como $100 - 65 = 35$, esto quiere decir que ¡se eliminará por error un 35 % de todos los individuos que den positivo en el test! El elevado porcentaje desaconseja totalmente el uso del test en este caso.

Nota: La situación que acabamos de describir no es tan particular como parece. De hecho, es bastante común, de ahí su interés (al margen de la mención anecdótica de la novela de Dick). En general, se podría enunciar así:

Tenemos una población "grande" de individuos (u objetos). De ellos, algunos poseen una característica que nos interesa especialmente, digamos A. La falta de la característica se denota por H. Todos los individuos son, por tanto, A o H, y no los podemos distinguir a simple vista. Podemos pensar que la característica A corresponde a ser androide y H a ser humano, como en el caso de Blade Runner, pero también podrían ser "tener cierta enfermedad" y "no tenerla", o "tener cierto defecto" y "no tener el defecto", por ejemplo.

Existe una prueba fiable para clasificar a los individuos (u objetos) como poseedores o no de la característica A, pero es cara, costosa y/o peligrosa. Afortunadamente existe un test mucho más rápido y barato para la clasificación. Si el test da positivo $(+)$ el individuo se clasifica como poseedor de la característica A y se clasifica como no poseedor de la característica A (como H) si da negativo $(-)$. El problema es que el test no es completamente fiable: puede fallar dando falsos positivos o falsos negativos. Sean

$$\alpha = P(+/H) \text{ la probabilidad de falso positivo y}$$

$$\beta = P(-/A) \text{ la probabilidad de falso negativo.}$$

Ambas son características del test, que se conocen o se estiman previamente a instaurar el uso generalizado del test en toda la población. A partir de $p = P(A)$, la proporción de individuos de la clase A en la población (proporción de androides, de enfermos, o de objetos con defectos, en los ejemplos mencionados), que es una característica de la población que se considere, se calcula el **valor predictivo positivo** o probabilidad de que un individuo (u objeto) realmente sea de la clase A si ha dado positivo, mediante la **fórmula de Bayes** como

$$P(A/+) = \frac{P(+/A)\,P(A)}{P(+/A)\,P(A) + P(+/H)\,P(H)} = \frac{(1-\beta)\,p}{(1-\beta)\,p + \alpha\,(1-p)}$$

y depende de las características del test, α y β, y de la población en la que se aplique el test, a través del valor de p. Si el valor predictivo positivo no es grande, se desaconseja el uso del test en esa población.

¿Y el **valor predictivo negativo**? También es importante este valor, que es la probabilidad de que el individuo sea de la clase H si ha dado negativo en el test (proporción de la clase H de entre los que dan negativo, $P(H/-)$), y también interesa que sea grande. La fórmula para calcularlo, análogamente a la del valor predictivo positivo es, usando la **fórmula de Bayes**,

$$P(H/-) = \frac{P(-/H)\,P(H)}{P(-/H)\,P(H) + P(-/A)\,P(A)} = \frac{(1-\alpha)\,(1-p)}{(1-\alpha)\,(1-p) + \beta\,p}$$

Para el ejemplo de los androides, recordemos que $\alpha = 0{,}01$ y $\beta = 0{,}06$. En la primera población teníamos que $p \cong 0{,}166667$. Entonces, el valor predictivo negativo sería

$$P(H/-) \cong \frac{(1-0{,}01)(1-0{,}166667)}{(1-0{,}01)(1-0{,}166667)+0{,}06\times 0{,}166667} = \frac{0{,}82499967}{0{,}83499969} = 0{,}9880239237$$

Para la segunda población, $p \cong 0{,}0196078$ y el valor predictivo negativo es

$$P(H/-) \cong \frac{(1-0{,}01)(1-0{,}0196078)}{(1-0{,}01)(1-0{,}0196078)+0{,}06\times 0{,}0196078} \cong \frac{0{,}970588}{0{,}97176475} = 0{,}9987890588$$

Notemos que ambos valores son bastante altos y que es mayor el de la segunda población, que es justamente la tiene menor valor predictivo positivo.

Queda claro en las anteriores fórmulas que el valor de p (que depende de la población) interviene en el cálculo de los valores predictivos del test. Éstos no pueden, por tanto, ser usados como índices para comparar dos tests o pruebas diagnósticas diferentes, salvo que se vayan a usar exactamente sobre la misma población. Por ello resulta útil determinar otros índices de valoración de un test que no dependan de p (esto es, que no dependan de la población).

Así, además de los conceptos de *sensibilidad* y *especificidad* de un test, se suele hablar de las *razones de probabilidad*, que miden cuánto más probable es un resultado concreto (positivo o negativo) según la presencia o ausencia de la característica A. Concretamente, hay dos razones de probabilidad, la *razón de probabilidad positiva* y la *razón de probabilidad negativa* que son, respectivamente,

$$RP_+ = \frac{1-\beta}{\alpha} = \frac{P(+/A)}{P(+/H)} \quad \text{y} \quad RP_- = \frac{1-\alpha}{\beta} = \frac{P(-/H)}{P(-/A)}\,.$$

En el ejemplo, como $\alpha = 0{,}01$ y $\beta = 0{,}06$, tenemos que

$$RP_+ = \frac{0{,}94}{0{,}01} = 94 \quad \text{y} \quad RP_- = \frac{0{,}99}{0{,}06} = 16{,}5$$

que se interpretan así: aproximadamente, es 94 veces más probable que el test dé + en androides que en humanos, y es 16,5 veces más probable que dé − en humanos que en androides, respectivamente.

Rosario Delgado
Dept. de Matemàtiques
Universitat Autònoma de Barcelona
08193 Bellaterra
delgado@mat.uab.cat

Publicat el 18 de setembre de 2006