ReaR
ISNN 1989 4090
Revista electrónica de AnestesiaR
Abril 2024

ORIGINAL

Un intruso de otro mundo: F1-score.

Molina Arias M

Hospital Infantil Universitario La Paz, de Madrid.

Resumen

El F1-score, también llamado F-score o medida F, es un estimador de la capacidad de clasificación de una prueba que se usa con frecuencia en la ciencia de datos y en los algoritmos de inteligencia artificial y que puede ser de utilidad para la valoración de las pruebas diagnósticas. Es la media armónica de sensibilidad y valor predictivo positivo, por lo que pondera el valor de ambos en un solo estimador.

Introducción

El F1-score, también llamado F-score o medida F, es un estimador de la capacidad de clasificación de una prueba que se usa con frecuencia en la ciencia de datos y en los algoritmos de inteligencia artificial y que puede ser de utilidad para la valoración de las pruebas diagnósticas. Es la media armónica de sensibilidad y valor predictivo positivo, por lo que pondera el valor de ambos en un solo estimador.

¿Alguna vez habéis sentido que un forastero ha irrumpido en vuestro mundo, un intruso que, aunque ajeno en apariencia, parece destinado a dejarse ver con cierta frecuencia? La ciencia de datos, un reino en constante expansión, ha traído consigo una nueva herramienta, el F1-score (la medida F1, para los puristas de la lengua de Cervantes), desconocida para muchos, al menos entre los adictos al mundo de la Medicina. Aunque pueda sonar como algo sacado de una película de ciencia ficción y nos parezca tan misterioso como un ovni, el F1-score ha decidido aterrizar en el terreno de las publicaciones médicas, donde parece haber encontrado un inesperado hogar.

Así que, si estáis listos para descubrir cómo este intruso venido de otro mundo puede mejorar la forma en que valoramos nuestras pruebas diagnósticas, acompañadme en este viaje por el sorprendente cruce entre la ciencia de datos y la Medicina.

Nuestras métricas habituales

Aunque existen múltiples herramientas para la valoración de las pruebas diagnósticas, las cuatro más queridas por el público son la pareja formada por sensibilidad y especificidad, y los dos valores predictivos, el positivo y el negativo.

Recordemos en qué consisten.

Cuando queremos evaluar una prueba diagnóstica, lo habitual es compararla con otra prueba que consideramos el patrón de referencia, lo que los amantes del inglés denominan el *gold standard*. Los positivos y negativos de las dos pruebas se representan en una tabla de

contingencia y, utilizando los valores de las casillas, hacemos nuestros cálculos.

Podéis ver, en la tabla adjunta, un ejemplo ficticio que compara dos pruebas para diagnosticar esa terrible enfermedad que es la fildulastrosis. Por un lado, la fildulastrina magnética (FM), nuestro patrón de referencia. Por otro, una prueba nueva, pero muy prometedora, el glóbulo verde (GV).

	Fildulastrina magnética (patrón de referencia)		
Glóbulo verde	POS	NEG	
POS	92 (VP)	85 (FP)	177 (GV POS)
NEG	8 (FN)	1815 (VN)	1823 (GV NEG)
	100 (Enfermos)	1900 (Sanos)	2000 (Total)

$S = \frac{VP}{(VP+FN)} = \frac{92}{(92+8)} = 0{,}92$ $\quad$ $VPP = \frac{VP}{(VP+FP)} = \frac{92}{(92+85)} = 0{,}52$

$E = \frac{VN}{(FP+VN)} = \frac{1815}{(85+1815)} = 0{,}95$ $\quad$ $VPN = \frac{VN}{(VN+FN)} = \frac{1815}{(1815+8)} = 0{,}99$

Tabla de contingencia que compara los resultados de una prueba diagnóstica con el estándar de referencia.

Podemos ver que nuestra muestra está compuesta por 2000 sujetos, 100 de los cuales padecen fildulastrosis y 1900 sanos.

De los 100 enfermos, 92 tienen una prueba de GV positiva. Son los verdaderos positivos (VP). Además, de los 1900 sanos, 1815 tienen GV negativo. Son los verdaderos negativos.

Pero vemos que la prueba del GV se confunde con algunos. Ocho de los enfermos dan negativo (falsos negativos, FN) y 85 de los sanos dan positivos (falsos positivos, FP). Con estas cuatro casillas construimos nuestros indicadores para valorar la prueba en estudio (el GV en nuestro ejemplo).

Sensibilidad y especificidad

La sensibilidad (S) es la capacidad para clasificar correctamente a los enfermos. Es el cociente entre los VP y el total de enfermos (VP + FN). En nuestro ejemplo, 0,92.

La especificidad (E) es la capacidad para clasificar correctamente a los sanos. Es el cociente entre los VN y el total de sanos (FP + VN). En nuestro ejemplo, 0,95.

Valores predictivos

Vamos con los valores predictivos. El valor predictivo positivo (VPP) es la proporción de positivos que están enfermos. Es el cociente entre los VP y todos los positivos (VP + FP). En nuestro ejemplo, 0,52.

El valor predictivo negativo (VPN) es la proporción de negativos que están sanos. Es el cociente entre VN y todos los negativos (VN + FN). En nuestro ejemplo, 0,99.

Para completar nuestro análisis de la tabla, calculemos la prevalencia de enfermedad. Sabemos que es el total de enfermos dividido entre el total de participantes. En nuestro ejemplo, 0,05 (o 5%, como prefiramos).

El problema de la falta de equilibrio

Vista la tabla de ejemplo y los cálculos que hemos hecho, ¿os parece que la prueba del GV es una prueba diagnóstica potente?

Los más aplicados me diréis que eso es difícil de contestar sin conocer los cocientes de probabilidad de la prueba (también llamados cocientes de verosimilitud) y, sin duda, tendréis razón.

Vamos a calcularlos.

El cociente de probabilidad positivo (CPP) nos dice cuánto más probable es tener un positivo en un enfermo que en un sano. Sabemos que la probabilidad de dar positivo en un enfermo es la S y que la probabilidad de dar positivo en un sano será la de clasificar

erróneamente a ese sano, o sea, el complementario de la E. Si calculamos S / (1 – E) = 18,4.

El cociente de probabilidad negativo (CPN) nos dice cuánto más probable es encontrar un negativo en un enfermo que en un sano. La probabilidad de encontrar un negativo en un enfermo será la de no clasificarlo correctamente, o sea, el complementario de la S. La probabilidad de que un sano dé negativo es la E. Si calculamos (1 – S) / E = 0,05.

Un CPP > 10 indica que la prueba es muy potente para el diagnóstico cuando da un resultado positivo.

De igual manera, un CPP < 0,1 también nos dice que la prueba es muy potente para descartar la enfermedad cuando da negativo.

Sin embargo, vemos que, aunque S y E tienen unos valores muy buenos, el VPP es bastante mediocre (0,52). ¿A qué se debe esto?

Efectivamente, el culpable de este valor predictivo tan bajo es la prevalencia de la enfermedad, que es baja. Para un mismo valor de S y E (o de cocientes de probabilidades), el VPP disminuye al ser menor la prevalencia de la enfermedad. En nuestro ejemplo, además, esta dificultad para clasificar correctamente los pacientes se ve agravada por la diferencia tan grande entre las proporciones de enfermos y sanos.

Pensad que si, en vez de hacer la prueba diagnóstica, siempre decimos que la persona está sana, acertaremos un 95% de las veces. La prueba no tiene una tarea nada sencilla, pero son cosas del azar y del teorema de Bayes.

El regateo diagnóstico

Llegados a este punto, si queremos valorar la utilidad de la prueba para ayudarnos a determinar si un paciente concreto tiene o no sufre la enfermedad, tendremos que calibrar, fundamentalmente, su S y su VPP.

La S nos dice la probabilidad de que un enfermo dé un resultado positivo, pero una vez que ya sabemos que está enfermo. Visto de otro modo, nos da una idea de los enfermos que vamos a ser capaces de diagnosticar con la prueba. Si la S es muy alta, habrá pocos enfermos a los que hagamos la prueba que queden sin un diagnóstico positivo.

El VPP no dice una cosa bien diferente: la probabilidad de que un positivo esté enfermo. Si el VPP es bajo, habrá personas sanas que se diagnosticarán como enfermos (FP), tantos más cuanto menor sea el VPP.

El problema es que, en muchas situaciones, especialmente cuando las distribuciones de los valores de la prueba entre sanos y enfermos no están bien separadas, cuando uno de los dos mejore, el otro empeorará, y viceversa.

¿Cuál nos interesa más?

Si la enfermedad es muy grave, nos interesará una S elevada para que no se nos escape ningún enfermo. El precio que habrá que pagar será un número más o menos elevado de FP.

Por el contrario, imaginad una enfermedad no tan grave y cuyo tratamiento sea caro o molesto. Preferiremos no tener FP, aunque se nos escapen algunos enfermos sin diagnosticar. En este caso, nos interesará tener mejor VPP, aunque la S sea menor.

En cualquier caso, nos vendría tener bien un único parámetro que nos resumiese el comportamiento global de la prueba en cuanto a S y VPP, sobre todo si estamos tratando de elegir cuál puede ser más útil entre diferentes opciones.

Aquí es donde nuestro intruso de otro mundo viene en nuestro auxilio.

Un intruso acude en nuestro auxilio: F1-score

El F1-score, también llamado F-score o medida F, es un estimador de la capacidad de clasificación de una prueba que se usa con frecuencia en la ciencia de datos y en los algoritmos de inteligencia artificial y que, últimamente, se va abriendo paso entre los trabajos de Medicina.

El F1-score es la media armónica de la S y el VPP, por lo que podemos definirlo según la siguiente fórmula:

$F1 = 2 / (S^{-1} + VPP^{-1})$

Esta fórmula suele transformarse en su versión más amigable, que es la siguiente:

$F1 = 2 \times S \times VPP / (S + VPP)$

Los valores posibles del F1-score oscilan entre 0 y 1. Una prueba perfecta (un clasificador perfecto, como se diría en la ciencia de datos) tiene un F1-score = 1 (tanto su S como su VPP valdrán 1). En el otro extremo, el valor mínimo posible es 0, que se producirá cuando S y/o VPP valgan 0.

De esta forma, el F1-score da una idea global del desempeño de la prueba en función de su S y su VPP. En nuestro ejemplo, el F1-score = 0,65, que nos indicaría, en principio, que la prueba tiene una moderada capacidad para discriminar sanos y enfermos (lo que ya nos anunciaban sus cocientes de probabilidad).

Imaginemos que el resultado de nuestra prueba del glóbulo verde es un valor continuo y que nosotros tenemos que definir el punto de corte para distinguir entre positivos y negativos. En este caso, podemos aumentar la S bajando el punto de corte, pero tendremos muchos falsos positivos (el VPP será menor).

Por el contrario, si aumentamos el punto de corte, mejoraremos el VPP de la prueba, pero probablemente se nos empezarán a escapar pacientes sin diagnosticar (bajará la S).

Podemos utilizar el valor del F1-score según evaluamos los diferentes puntos de corte. Por ejemplo, si nos interesa aumentar el VPP podremos ir aumentando el punto de corte hasta el momento en que el valor del F1-score empiece a bajar de forma llamativa. Esto querrá decir que, probablemente, habremos sacrificado en exceso la S de la prueba en nuestro empeño por mejorar su VPP, con lo que el número de pacientes que queden sin diagnosticar puede que sea más elevado de lo que nos convenga (eso sí, el número de falsos positivos será menor).

Una cosa que debemos tener en cuenta es que el F1-score, al depender directamente del VPP, comparte con él el defecto de depender de la prevalencia de la enfermedad. Como es lógico, una misma prueba diagnóstica ensayada en dos poblaciones distintas mostrará un valor de F1-score mayor en la población en la que la prevalencia de la enfermedad sea más alta.

Por este motivo, si queremos comparar pruebas entre poblaciones diferentes, quizás sea mejor recurrir a otros estimadores que no afecten (tanto) por la prevalencia, como los cocientes de

probabilidades o el área bajo las curvas ROC.

El intruso tiene familia

Hasta ahora hemos hablado de F1-score pero, en realidad, sería más correcto hablar del F-score (sin el 1) cuando nos referimos al estimador de forma general.

Ya hemos visto que F-score representa un balance entre S y VPP. La situación más habitual es un balance equilibrado entre los dos parámetros, pero puede haber ocasiones en que nos interese dar prioridad al alguno sobre el otro.

Así, nos encontramos con toda una familia de medidas Fβ, siendo β el parámetro que nos permite escoger el balance entre S y VPP que más nos interese.

Así, podemos entender el Fβ-score como una abstracción del F-score en la que el cálculo de la media armónica de S y VPP se controla por este parámetro β. Podemos ver cómo quedaría la ecuación para el cálculo:

$F\beta = ((1 + \beta^2) \times S \times VPP) / (\beta^2 \times (S + VPP))$

El valor neutro en este balance es el que se corresponde con β = 1. En ese caso, la ecuación anterior queda como la media armónica sin modificar y obtenemos la medida F que pondera equilibradamente S y VPP.

Aunque, en teoría, podríamos elegir el valor de β que quisiésemos, en la práctica suelen utilizarse solo tres de ellos:

- β = 0,5. Da más importancia al VPP que a la S, por lo que nos ayudará a minimizar el número de falsos positivos. Lo usaremos para establecer el punto de corte cuando nos perjudique más tener falsos positivos que falsos negativos (que se nos escapen enfermos sin diagnosticar).
- β = 1. Es el F1-score del que hemos hablado antes. Nos balancea de forma similar la S y el VPP (o los falsos positivos y negativos).
- β = 2. Este valor disminuye el peso del VPP y aumenta el de la S. O sea, se prefiere minimizar los falsos negativos (pacientes sin diagnosticar) aunque aumenten los falsos positivos.

Como veis, es un regateo continuo. En esto de las pruebas diagnósticas, como en la vida, no siempre se puede tener todo y hay que elegir qué preferimos priorizar.

Nos vamos…

Ya veis que el mundo de las pruebas diagnósticas es amplio y que en él caben muchos estimadores diferentes para valorar la capacidad de desempeño de las pruebas.

Esto es debido a que ninguna prueba es perfecta en todo, por lo que la mayor parte de las veces tendremos que elegir si favorecemos los falsos positivos, los falsos negativos o lo que nos interese más.

Hemos mencionado, aunque muy de pasada, que este problema puede acrecentarse cuando hay mucho desequilibrio entre la proporción de enfermos y sanos (prevalencia de enfermedad muy baja). En estos casos, la tarea de la prueba diagnóstica se complica y puede costar elegir el punto de corte más adecuado a nuestras necesidades.

Contamos, además de con el F-score, con alguna otra medida, también venida del mundo de la ciencia de datos, que

nos ayuda con el regateo entre S y VPP. Me refiero, concretamente al denominado enriquecimiento de la precisión respecto a la recuperación (o del VPP respecto a la S, en nuestro lenguaje habitual), muy relacionado con los conceptos de probabilidad preprueba y postprueba.

Pero esa es otra historia…

Bibliografía

- *Assessment of diagnostic tests. En: Palmas WR, ed. Pocket evidence based medicine. A survival guide for clinicians and students*. Springer. NY, 2023; 15-33. (HTML)
- *Linear and logistic regression*. En: Zumel N, Mount J, eds. Practical Data Science with R, 2ª ed. Manning Publications Co. Shelter Island, NY, 2020;215-73. (HTML)
- Sasaki Y. *The truth of the F-measure.* School of Computer Science, University of Manchester, 2007. [Disponible en: https://www.researchgate.net/publication/268185911_The_truth_of_the_F-measure]. [Consultado en 12/09/2023].

Correspondencia al autor

Manuel Molina Arias
mma1961@gmail.com
Servicio de Gastroenterología
Hospital Infantil Universitario La Paz, de Madrid.

Aceptado para el blog en noviembre de 2023