ReaR
ISNN 1989 4090
Revista electrónica de AnestesiaR
Enero 2024

ORIGINAL

El decálogo inteligente. Lectura crítica de trabajos que emplean aprendizaje automático.

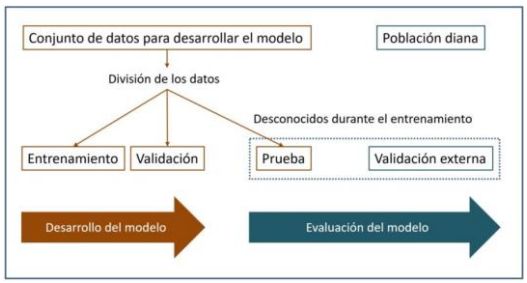
Molina Arias M.

Hospital Infantil Universitario La Paz, de Madrid.

Resumen

Se revisan los aspectos que deben valorarse para la lectura crítica de trabajos que emplean técnicas de aprendizaje automático, incluyendo la selección de participantes, el tratamiento de los datos durante el desarrollo del modelo y su implementación final en la práctica clínica.

Introducción



Se revisan los aspectos que deben valorarse para la lectura crítica de trabajos que emplean técnicas de aprendizaje automático, incluyendo la selección de participantes, el tratamiento de los datos durante el desarrollo del modelo y su implementación final en la práctica clínica.

Hoy vamos a pasearnos por el fascinante y divertido mundo de los decálogos.

Ya sabéis, esas listas de reglas y consejos que nos rodean como moscas en una merienda campestre. Un buen ejemplo es el decálogo del buen escritor, culpable del sufrimiento de no pocos árboles por acción de muchos aspirantes a Nobel de Literatura.

Otro es el decálogo del conductor, que debería incluir una regla número once: «si ves un peatón distraído con la nariz en su móvil, álzalo gentilmente y devuélvelo a su hábitat natural, la acera».

O uno de mis favoritos, el decálogo del político honesto que dice… pero ¡esperad!, parece que he perdido esta lista. Tal vez esté con aquel calcetín que metí en la lavadora y que nunca volví a ver. Es una pena, seguro que era el más entretenido.

Hay tantos decálogos famosos por ahí como realidades alternativas en Matrix, así que no me he podido resistir y me he puesto a pensar en uno nuevo —el decálogo inteligente—, con el que reflexionar sobre la calidad de los métodos de inteligencia artificial y aprendizaje automático que podemos encontrarnos, cada vez con más frecuencia, en los trabajos científicos.

Así que, ahora que hemos ampliado nuestro horizonte decalógico, vamos a sumergirnos en las turbias aguas de la **lectura crítica de documentos científicos basados en inteligencia artificial,** donde los algoritmos y los datos clínicos van de la mano para descubrir los secretos ocultos detrás de la magia del aprendizaje automático.

Unas aclaraciones previas

Antes de pasar a desglosar los puntos de nuestro decálogo, vamos a centrar un poco el tema de los algoritmos y los métodos de aprendizaje automático.

Ya vimos en una entrada anterior la diferencia entre aprendizaje automático supervisado y no supervisado. En el primero, proporcionamos al algoritmo un conjunto de datos con las respuestas correctas ya conocidas para que aprenda a generalizar y hacer predicciones precisas sobre nuevos datos. Estos modelos, una vez entrenados, se utilizan para predecir el valor desconocido de una variable a partir de una serie de variables conocidas.

En el segundo, el algoritmo de aprendizaje automático no supervisado no recibe información etiquetada sobre las respuestas correctas. En lugar de eso, se basa en la estructura y relaciones presentes en los datos de entrada para descubrir patrones y características interesantes que son desconocidas para el investigador.

Lo cual nos lleva a preguntarnos qué es y cómo funciona un algoritmo.

Brevemente, podemos definir un algoritmo como un conjunto ordenado de instrucciones o reglas precisas que se utilizan para resolver un problema o llevar a cabo una tarea en un número finito de pasos. De manera esquemática, todos ellos responden a la siguiente ecuación:

$f(-;\theta)$: $x \rightarrow f(x; \theta) = y(\theta)$

Antes de que nadie se alarme, dejadme que os explique qué significa. En primer lugar, «f» es una función que hace unos cálculos específicos (que dependerán del algoritmo) empleando unos parámetros determinados (θ). De esta forma, nosotros introducimos nuestra variable conocida (x) en la función y esta hará los cálculos para obtener la variable que queremos predecir (y). De esta forma, el trabajo del algoritmo, que se desarrolla durante la fase de entrenamiento, es aprender qué valores de los parámetros permiten hacer una buena predicción de «y».

Una vez entrenado, el modelo (ya no le llamamos algoritmo) es capaz de predecir el valor de «y» a partir de valores de «x» que nunca había visto.

Estos modelos pueden ser muy potentes, pero para funcionar bien deben ser alimentados con datos de buena calidad. Una parte importante de todo el proceso es preparar los datos, lo que se llama ingeniería de características (*feature engineering*): manejo de valores faltantes y extremos, creación de variables nuevas, reducción de la dimensionalidad de los datos, etc.

Algunos de estos algoritmos funcionan mejor si los datos de todas las variables se encuentran en un rango de valores absolutos similar. Por eso, es frecuente que las variables se estandaricen o se normalicen, que no es exactamente lo mismo.

La normalización, o escalamiento lineal, consiste en transformar los valores para que todos se encuentren dentro de un rango determinado (habitualmente entre 0 y 1). Por su parte, para estandarizar restamos a cada valor la media y la dividimos entre la desviación estándar, de tal modo que los datos siguen una distribución de media 0 y desviación estándar 1.

Otro aspecto fundamental es que los autores del algoritmo hayan tomado las medidas necesarias para evitar que se haya producido un sobreajuste (*overfitting*) durante la fase de entrenamiento. Si esto se produce, el modelo lo que hace es aprenderse «de

memoria» patrones de los datos de entrenamiento que no tienen por qué ser comunes con los datos de la población general, desconocidos para el modelo.

Si se produce sobreajuste, el modelo funcionará muy bien sobre los datos de entrenamiento, pero no será capaz de generalizar sus predicciones cuando le ofrezcamos datos que no haya «visto» antes.

Este es el motivo de que sea obligado hacer una división del conjunto de datos en varios bloques: entrenamiento, validación y prueba, tal como veis en la figura.

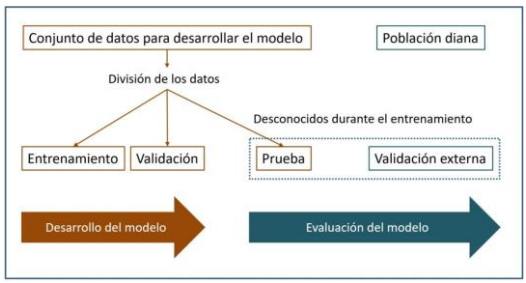


Figura. Separación de los datos en tres conjuntos: entrenamiento, validación y prueba.

Así, el modelo se entrena con los datos de entrenamiento, utilizando los datos de validación para tratar de controlar el sobreajuste durante el entrenamiento del modelo. Una vez entrenado, probaremos su desempeño con el conjunto de datos de prueba.

Y la cosa no acaba aquí. Como, con cualquier método estadístico más «convencional», quedaría una fase de validación externa en una población diferente a aquella de la que procedían los datos de entrenamiento, validación y prueba.

Hechas estas consideraciones generales sobre los algoritmos de aprendizaje automático y sus datos, vayamos con el decálogo prometido.

El decálogo inteligente

Antes de ir con el decálogo, permitidme que os dé dos consejos para valorar mejor este tipo de trabajos.

El primero, no os olvidéis de lo básico.

Los trabajos que emplean técnicas de inteligencia artificial o aprendizaje automático suelen desarrollar pruebas diagnósticas o reglas de predicción clínica. Valorad siempre la metodología del trabajo siguiendo las directrices para la valoración crítica de las pruebas diagnósticas o de las reglas de predicción.

El segundo, muy relacionado con el anterior, **no os dejéis impresionar por la metodología,** por muy puntera y compleja que pueda parecer.

En el fondo, estos algoritmos no dejan de ser técnicas estadísticas posibilitadas por la mayor potencia de cálculo y de manejo de datos. Una red neuronal compleja puede esconder un trabajo de metodología deficiente.

Para eso está la lectura crítica.

Ahora sí, veamos 10 reglas que deberían cumplirse durante el desarrollo y la aplicación de los métodos de aprendizaje automático en los trabajos científicos.

1. Debe obtenerse una muestra suficiente de participantes adecuados.

La obtención de los participantes y el tamaño muestral necesario dependerán del contexto específico del estudio.

Los participantes deben ser representativos de la población en la que se quiere aplicar el modelo en la práctica. Un modelo puede ser perfecto desde el punto de vista metodológico y resultarnos inservible porque la muestra

en la que se obtuvo sea muy diferente de la población de nuestro medio.

El tamaño muestral también dependerá de las características del estudio. En algunas técnicas de aprendizaje supervisado (regresión y clasificación) suele estar condicionado por el número de las variables utilizadas y la complejidad del modelo.

En otros casos de aprendizaje no supervisado, el tamaño muestral puede estar más en relación con la representatividad de los datos y la capacidad del modelo para extraer patrones significativos.

2. Definir previamente en qué contexto clínico quiere utilizarse el modelo.

Los modelos pueden emplearse como herramientas diagnósticas para detección y clasificación de enfermedades, determinación de factores pronósticos y personalización de tratamientos, entre otras posibilidades.

Es importante que se hayan tenido en cuenta los aspectos éticos sobre seguridad y confidencialidad de los datos de los pacientes, así como la accesibilidad de estos métodos por parte de los médicos encargados del cuidado de los pacientes.

3. La muestra de participantes debe representar todo el espectro de la enfermedad en la población diana.

Esto es común a cualquier estudio de prueba diagnóstica. Si en el conjunto de datos hay una sobrerrepresentación de casos más graves, el modelo incurrirá en un alto número de falsos positivos cuando lo apliquemos en la población diana, lo que se conoce como **sesgo de espectro**.

También en estos casos pueden producirse **sesgos de selección**, cuando la muestra de participantes no es representativa de todo el espectro de enfermos de la población.

4. Debe especificarse con claridad cómo se obtuvieron los datos de entrenamiento, validación y prueba.

Ya hemos hablado sobre la importancia de tratar los datos de forma correcta durante la fase de entrenamiento del modelo. Si los autores del trabajo no mencionan el proceso, por ejemplo, cómo se obtuvo el conjunto de validación, podremos sospechar que hayan podido usar los mismos datos para la validación interna y externa del modelo.

En algunas ocasiones, los datos pueden no ser muy numerosos y no es posible dividirlos en tres conjuntos separados. En estos casos, pueden emplearse técnicas de validación cruzada, que consisten en dividir los datos de entrenamiento en bloques y utilizar, de forma secuencial, uno de los bloques como conjunto de validación y, el resto, como conjunto de entrenamiento. Esto ayuda a prevenir los sesgos de selección y a controlar el sobreajuste del modelo.

5. El patrón de referencia desde ser adecuado y estar bien definido.

Una vez más, esto es común a cualquier estudio sobre pruebas diagnósticas, pero, en el caso de los modelos de aprendizaje supervisado, tenemos que prestar atención al modo en que se realizó el etiquetado de los datos en la fase de entrenamiento.

Si el patrón de referencia es un juicio clínico sobre una imagen o una combinación de variables, debe especificarse el grado de cualificación de los que hacen el diagnóstico, la variabilidad intra e interobservador y el

método usado para llegar al acuerdo en el diagnóstico.

6. Los resultados deben expresarse con la métrica adecuada.

Esto es común a cualquier documento científico. Dado que la mayoría son estudios de diagnóstico, pediremos a los autores que nos muestren la tabla de contingencia, que muchas veces recibe el nombre de «matriz de confusión» en el argot de la ciencia de datos.

Además, no nos conformaremos solo con términos como exactitud o precisión, muy queridos en este ambiente. A nosotros nos gustan las curvas ROC y los cocientes de probabilidades, que son los que nos permiten valorar de forma más adecuada la capacidad de una prueba diagnóstica.

Como siempre, si no nos los proporcionan los autores, intentaremos calcularlas nosotros mismos a partir de los datos del estudio.

7. Es aconsejable que el clínico comprenda el modelo.

En muchas ocasiones, estos modelos funcionan como una caja negra en la que vemos entrar datos y salir resultados sin tener ni idea de qué ocurre dentro del modelo.

Debemos hacer un esfuerzo (y los autores deben facilitar la tarea) por comprender el modelo, aunque, por desgracia, suele haber una relación inversa entre la capacidad del modelo y el grado de compresión al que puede llegar el clínico.

8. Sospecha del modelo demasiado bueno.

Aunque sea producto de una superpoderosa inteligencia artificial, siempre sospecharemos del modelo que funcione llamativamente bien.

Seguro que los más atentos ya habréis imaginado quién puede ser el culpable de esto. Pues sí, el sobreajuste durante la fase de entrenamiento. Deberemos revisar si se ha tenido en cuenta y cómo ha tratado de controlarse durante la fase de desarrollo del modelo.

9. El trabajo debe ser reproducible.

Esta es una de las premisas del método científico. Lo ideal sería que los autores proporcionasen el código de programación del algoritmo y los datos de entrenamiento, con lo que podríamos reproducir el proceso nosotros mismos.

Sin embargo, esto no suele ser habitual en el mundo de competencia y secretismo en los que se está introduciendo todo el desarrollo de la inteligencia artificial. Si tenemos los conocimientos necesarios (que no suele ser el caso entre los clínicos), siempre podemos escribir el código de manera independiente y reproducir los resultados.

10. El modelo debe ser útil para el paciente.

Esto es lógico, el paciente debe ser el beneficiario final de todo el proceso de desarrollo e implementación del modelo. Si no va a modificar el tratamiento o el pronóstico del paciente, será un modelo inútil por muy bien desarrollado que esté.

Una propuesta de lista de verificación

Y hasta aquí nuestro decálogo. Para sacarle un poco más de utilidad, una vez reflexionemos sobre los distintos puntos referidos, podemos hacernos diez preguntas a las que contestar «sí», «no» o «no sé», a semejanza de las parrillas de lectura crítica de la red CASPe, de

las que no existe un ejemplo para métodos de aprendizaje automático.

Estas diez preguntas serían las siguientes:

- ¿Se hizo la selección de participantes de manera correcta?
- ¿Los datos eran de calidad y se trataron de forma adecuada (ingeniería, limpieza, valores faltantes, valores extremos…)?
- ¿Se hizo el entrenamiento, la validación y la prueba interna de forma adecuada?
- En el aprendizaje supervisado, ¿se realizó el etiquetado de forma correcta?
- ¿Incluyeron los datos todo el espectro de la enfermedad estudiada y con categorías bien balanceadas?
- ¿Se eligió el modelo correcto?
- ¿Se consideró el modelo más simple de los posibles?
- ¿Fue la muestra de entrenamiento representativa de la población diana?
- ¿Se interpretaron los resultados de manera correcta?
- ¿Se hizo una validación externa del modelo y se valoró su impacto clínico?

Una vez contestadas estas diez preguntas, estaremos en condiciones de valorar la metodología relacionada con la técnica de aprendizaje automático empleada.

Ya solo nos quedará hacer la valoración global del trabajo, para lo que utilizaremos cualquiera de las herramientas de lectura crítica disponibles en Internet.

Nos vamos…

Y aquí lo vamos a dejar por hoy.

Hemos visto, de manera general, la importancia del tratamiento de los datos para el desarrollo de modelos de aprendizaje automático y los parámetros de calidad del estudio que debemos valorar cuando hagamos una lectura crítica del mismo.

Ya hemos dicho que la mayor parte de estas técnicas se emplean, a día de hoy, para desarrollar herramientas diagnósticas. El problema que surge es que, en ocasiones, las personas que desarrollan los algoritmos provienen de la ciencia de datos y están más acostumbrados a utilizar métricas que a los clínicos nos resultan menos familiares.

Con el tiempo, es probable que tengamos que familiarizarnos con esta nueva jerga de matrices de confusión, precisión, puntuación F y otras cosas así. Pero esa es otra historia…

Bibliografía

- Al-Zaiti SS, Alghwiri AA, Hu X, Clermont G, Peace A, Macfarlane P, et al. *A clinician's guide to understanding and critically appraising machine learning studies: a checklist for Ruling Out Bias Using Standard Tools in Machine Learning (ROBUST-ML).* Eur Heart J Digit Health. 2022;3:125-40. (PubMed)
- Faes L, Liu X, Wagner SK, Fu DJ, Balaskas K, Sim DA, et al. *A clinician's guide to artificial intelligence: how to critically appraise machine learning studies.* Trans Vis Sci Tech. 2020;9:7. (PDF)
- Vinny PW, Garg R, Srivastava MVP, Lal V, Vishnu VY. *Critical appraisal of a machine learning paper: a guide for the neurologist.* Ann Indian Acad Neurol. 2021;24:481-9. (PubMed)

Correspondencia al autor

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

Aceptado para el blog en junio de 2023