TECHNO Review | Vol. 9, No. 2, 2020 | ISSN 2695-9933
International Technology Science and Society Review / Revista Internacional de Tecnología Ciencia y Sociedad
https://doi.org/10.37467/gka-revtechno.v9.2814

INTELIGENCIA ARTIFICIAL Y MEDICINA

La necesidad de modelos interpretables

Artificial Intelligence and Medicine: the Need for Interpretable Models

SARA LUMBRERAS

Universidad Pontificia Comillas, España

KEY WORDS

Artificial Intelligence
Healthcare
Automatic Diagnosis
Predictive Models
Black box
Interpretable AI
Covid-19

ABSTRACT

The pandemic has provided clear examples of the potential of AI for the health sector, as well as some of its issues, largely derived from the use of black box models. In some cases, there are no reasonable alternatives, as in image and speech processing. However, in many other instances it would be more profitable to try to focus the developments on Interpretable AI, which could be used more directly for the confirmation of knowledge or for the generation of new hypotheses that can be tested with subsequent experiments.

PALABRAS CLAVE

Inteligencia Artificial
Ciencias de la salud
Sistemas de Diagnóstico
Modelos Predictivos
Caja negra
IA Interpretable
Covid-19

RESUMEN

La pandemia ha proporcionado ejemplos claros del potencial de la IA para el sector de la salud, así como algunos de sus problemas, en buena parte derivados del uso de modelos de caja negra. En algunos casos, no existen alternativas razonables a los modelos de caja negra, como en tratamiento de imagen y voz. Sin embargo, en muchas otras situaciones resultaría más provechoso intentar centrar los desarrollos en la línea de la IA interpretable, que podrían ser aprovechados de manera directa para la confirmación de conocimiento o para la generación de hipótesis nuevas que puedan comprobarse con experimentos posteriores.

Recibido: 13/01/2020
Aceptado: 15/01/2020

GLOBAL KNOWLEDGE ACADEMICS

1. Una emergencia sin precedentes

La pandemia ha generado una emergencia sin precedentes que se ha materializado no sólo a nivel médico sino también científico: el conocimiento es necesario para apoyar la toma de decisiones y hemos necesitado generarlo a una velocidad mucho más rápida de la que estamos acostumbrados.

La inteligencia artificial (IA) ha demostrado ser una herramienta de valor incalculable para generar nuevo conocimiento y herramientas de utilidad inmediata. Además, la situación actual ha dejado patente que contar con datos fiables y transparentes es imperativo para poder apoyar la toma de decisiones técnicas y políticas.

La IA trabaja sobre los datos y construye modelos de manera mucho más rápida que los basados en la experiencia humana. Esta necesidad de velocidad se ha visto también reflejada en la tremenda aceleración que ha experimentado la generación de conocimiento. Tanto es así, que esta época ha sido denominada por algunos el "auge de los preprint", refiriéndose a los artículos científicos que se difunden, dada su urgencia, cuando aún no han superado el proceso de revisión por pares previo a su publicación.

La velocidad y el aprovechamiento de los datos según se generan, que tanto necesitamos en tiempos de crisis, son precisamente las principales ventajas de la IA. Este artículo repasa algunos de los desarrollos más interesantes en este contexto y extrae las lecciones más importantes sobre la IA y sus oportunidades y limitaciones en el contexto de las ciencias de la salud.

2. Nuevas aplicaciones de la IA en las ciencias de la salud

Una de las primeras aplicaciones de la IA en este contexto fue la creación de pruebas diagnósticas en los momentos en los que no existía disponibilidad de pruebas PCR. Se consiguió desarrollar alternativas a la PCR que funcionaban a partir de imágenes de tomografía de tórax con una alta fiabilidad de diagnóstico (Ardakani, Kanafi, Acharya, Khadem, & Mohammadi, 2020).

El rastreo de contactos ha sido otra área en la que la IA ha realizado aportaciones significativas. Más de 36 países han empleado herramientas digitales para el rastreo de contactos con relativo éxito (Lalmuanawma, Hussain, & Chhakchhuak, 2020). En estas aplicaciones, la IA complementa el rastreo manual, en ocasiones introduciendo elementos de teoría de grafos.

La IA nos ha ayudado también en la búsqueda de tratamientos efectivos para la Covid. Desde hace tiempo se utilizan las técnicas de simulación molecular para identificar fármacos que podrían tener interacciones con una sustancia dada (por ejemplo, con las proteínas de la espícula del virus). Las técnicas de simulación ayudaron a identificar moléculas como el Remdisivir y el Antazanavir como posibles fármacos para la lucha contra la Covid (Beck, Shin, Choi, Park, & Kang, 2020). Estos fármacos no se aceptan de manera directa, sino que se someten a ensayos clínicos específicos. La IA nos ayuda, así, a centrar los esfuerzos en opciones interesantes sin dejar de cumplir los procedimientos establecidos.

El análisis de los síntomas para predecir la evolución de los enfermos ha sido otro contexto en el que las aportaciones de la IA han resultado insustituibles. Por ejemplo, en una serie de proyectos en los que he tenido el placer de participar, la IA ha identificado los principales factores de riesgo que incrementan la necesidad de ingreso en cuidados intensivos (Izquierdo, Ancochea, Soriano, & Savana COVID-19 Research Group, 2020), las diferencias que manifiesta la enfermedad entre hombres y mujeres (Ancochea, Izquierdo, Savana COVID-19 Research Group, & Soriano, 2020) o las interacciones entre la Covid y la enfermedad pulmonar obstructiva crónica (Graziani et al., 2020).

También han emergido estudios en los que el análisis de datos ha arrojado conclusiones contradictorias o confusas, como es el caso de las asociaciones entre la gravedad de la enfermedad de Covid y el grupo sanguíneo de los pacientes (Zietz & Tatonetti, 2020).

En todas estas aplicaciones -y otras muchas- queda patente el potencial de la IA y el análisis de datos para apoyar el desarrollo de conocimiento nuevo de manera rápida,

aprovechando su capacidad de desvelar los patrones que aparecen en los datos.

3. ¿Puede la IA generar conocimiento médico?

Como decíamos, la IA es capaz de identificar en los datos patrones que ya sean conocidos por la comunidad científica (con lo cual, pueden confirmar este conocimiento) o desvelar patrones nuevos, como ha venido sucediendo en incontables aplicaciones de diagnóstico automático o de modelos predictivos aplicados a la salud.

Sin embargo, es necesaria la prudencia al hablar de generación de nuevo conocimiento: la IA puede como mucho, desvelar correlaciones, y corresponde a los expertos en cada problema particular establecer si estas correlaciones tienen como base un fenómeno clínico conocido, si responden a un fenómeno previamente desconocido pero razonable y que puede estudiarse a través de experimentos posteriores o si se deben a defectos de los datos de entrada (como podría ser el caso, por ejemplo, en el estudio sobre Covid y grupo sanguíneo). Estas tres situaciones diferentes aplicarían tanto a sistemas de diagnóstico como a modelos predictivos y aplicaciones de gestión.

4. Las cajas negras

Establecer en cuál de esos tres supuestos anteriores nos encontramos no es posible en el caso de los algoritmos denominados, de caja negra, en los que estas correlaciones no se hacen explícitas: el algoritmo (por ejemplo, una red neuronal) se entrena a partir de unos datos de entrada y devuelve una predicción para cada nueva instancia que le sea presentada. Sin embargo, no incluye ninguna explicación de la predicción que pueda utilizarse como punto de partida para comprenderla.

Esto resulta problemático en todo contexto en el que las decisiones que se tomen basadas en los modelos revistan alguna importancia, como es el caso comúnmente en las aplicaciones médicas. Para poder utilizar una herramienta, y mucho más para delegar una decisión, hemos de poder confiar en lo que nos comunica. ¿Cómo podríamos confiar en lo que no podemos comprender? Me refiero aquí a comprender el resultado complejo de la aplicación del código, aunque los mecanismos en esencia sean extremadamente simples. El resultado de la IA es complejo en el sentido en el que las ciencias de la complejidad aplican esta palabra (Mitchell, 2009): un sistema complejo es un sistema donde unas pocas reglas simples e interacciones de las partes dan lugar a fenómenos que no podemos deducir a partir de estas reglas. Esta emergencia, en el sentido de Chalmers (precisamente, la aparición de fenómenos inesperados (Chalmers, 2006)) hace que sea imposible anticipar el resultado del código a partir del mismo.

Sin embargo, los algoritmos de caja negra sólo nos dejan examinar eso: su código. Y esto presenta graves problemas. Si no sabemos cómo el algoritmo determina una decisión, ¿cómo podremos confiar en ella? Muchos, precisamente, desconfían de la IA por este motivo. Creo que esta desconfianza está plenamente justificada. Dos motivos resultan, en la práctica, los causantes de la mayoría de los problemas relacionados con la aplicación de sistemas de control delegado: el sobreajuste y el sesgo algorítmico.

5. Sobreajuste y sesgo algorítmico: dos problemas que la medicina no puede tolerar

En el sobreajuste, los datos que se le proporcionan a la máquina no son suficientes como para poder generalizar. Hemos de recordar que el aprendizaje automático extrae patrones en los datos que podríamos asemejar a "ejemplos de problemas resueltos" y que después aplica esas reglas a nuevas instancias del problema. Si los ejemplos que se le han proporcionado son suficientemente diversos y similares a los problemas a los que se aplicará después, el algoritmo probablemente funcionará bien. Sin embargo, en muchas ocasiones el algoritmo recibe menos datos de los que serían necesarios, con lo que, como un mal estudiante, acaba aprendiéndose los ejemplos de memoria, lo cual produce consecuencias desastrosas cuando intentamos generalizar. Si diseñamos un algoritmo para distinguir gatos de perros, pero todos los perros que aparecen en los datos de

entrenamiento son blancos y los gatos son negros, el algoritmo probablemente inferirá que los animales blancos se llaman perros y los negros se llaman gatos. Si se le muestran fotografías que no correspondan a estos colores, generalizará de manera desastrosa. Si pudiéramos acceder a las reglas que ha derivado el algoritmo sería posible auditarlas, pero esto no está dentro de nuestras posibilidades si trabajamos con una caja negra. Aún así, existen técnicas más o menos sofisticadas para intentar detectar posibles problemas de sobreajuste, pero por definición son difíciles de detectar.

De la misma manera, es posible que los sesgos presentes en los datos hagan que el algoritmo se base en variables que consideraríamos inadecuadas, lo que identificaríamos con el problema de sesgo algorítmico. Es bien conocido el ejemplo de ProPublica, la empresa que desarrolló un algoritmo de tipo caja negra para predecir la reincidencia criminal en los presos estadounidenses. Este algoritmo se lleva aplicando durante años para decidir si a los presos se les concede la libertad condicional. Tras análisis detallados de sus predicciones, se vio que una de las variables principales empleadas por el algoritmo era la raza: los presos afroamericanos, independientemente de su historial, recibían predicciones peores que los blancos. Esto se debía, al parecer, a la base de datos que se había empleado para entrenar al algoritmo: los presos de color de esos datos tenían una peor tasa de reincidencia. Otro ejemplo reciente, también relacionado con la raza, es el que emergió en una red social que filtraba con mucha más frecuencia las fotos de mujeres de color como contenido inapropiado. Al parecer, el algoritmo se entrenaba con fotografías publicitarias, por una parte, y con contenido pornográfico por la otra. Este último era mucho más racialmente diverso que la publicidad, lo que llevaba al algoritmo a asumir con más facilidad que la foto de una mujer que no fuese blanca correspondía a contenido inaceptable.

El sobreajuste y el sesgo algorítmico pueden ser extremadamente dañinos en cualquier contexto, pero en las ciencias médicas se vuelven intolerables. El nivel de fiabilidad que se le pide a los desarrollos médicos refleja la importancia de las decisiones que de ellos se derivan.

6. ¿Es posible superar las cajas negras?

Tanto el sobreajuste como el sesgo algorítmico podrían solventarse – o, al menos, detectarse, lo cual sería la primera etapa hacia su solución – si los algoritmos nos dejasen *ver* las reglas que infieren, si dejasen de ser cajas negras. Son muchos los que piensan que únicamente los modelos de caja negra alcanzan desempeños aceptables: se asume que sólo los modelos muy complejos, demasiado complejos como para poder comprenderse, son los que tienen alguna posibilidad de funcionar adecuadamente en los problemas reales. Sin embargo, esto no es cierto en todos los casos, ni siquiera en su mayoría. Pese a que en algunos casos (principalmente, el procesamiento de imagen o de sonido) sólo tienen aplicación los modelos especialmente complejos como las redes neuronales convolucionales y otras técnicas del espectro del aprendizaje profundo, en muchos otros problemas es posible desarrollar modelos transparentes. Cada vez recibe más apoyo la idea de la IA Interpretable (Molnar, 2020) . La hipótesis de base es que no existe un solo modelo que pueda realizar adecuadamente una tarea, sino que son posibles muchos modelos diferentes, con niveles de complejidad dispares. Algunos de ellos serán tan simples que sus reglas pueden expresarse de manera explícita (por tanto, no son cajas negras).

Por ejemplo, Cynthia Rudin desarrolló una alternativa a la caja negra "racista" de ProPublica (Rudin, 2019). El sistema era un esquema simple de puntuaciones en los que consideraba variables transparentes como la edad, el número de crímenes y de crímenes violentos cometidos en el pasado, con una puntuación determinada para cada posible valor. Este modelo tiene un poder predictivo muy similar a la caja negra de ProPublica, pero es completamente transparente. Los modelos basados en puntuaciones, las regresiones logísticas o los árboles de decisión son técnicas dominadas por todos los ingenieros que trabajen en este contexto, pero no son las más aplicadas porque generan una sensación de simplicidad (en el mal sentido), de falta de sofisticación cuando se las

compara con las cajas negras. Parece como si fuésemos un mago que no quisiera revelar sus trucos – o sus errores.

En los últimos trabajos en los que he aplicado técnicas de IA a un contexto de medicina he aprendido muchas cosas, pero una es la más importante: sólo es válido el modelo en el que podemos confiar, y sólo podemos confiar en algo que comprendemos y que además *tiene sentido* cuando lo ponemos en relación con la experiencia previa y el sentido común.

No puedo dar por válido un modelo que cuantifique el riesgo de que un paciente de Covid ingrese en UCI según la provincia que habita (¿o quizá sí?), o que prediga que padecer de diabetes disminuye el riesgo, o que los pacientes de 79 años tienen un riesgo decenas de veces mayor que los pacientes de 80 años y medio. Todos esos son resultados que pueden aparecer en un modelo, y que pueden ser matemáticamente correctos en el sentido de adaptarse adecuadamente a los datos que reciben. Sin embargo, no son válidos porque no se basan en reglas que tengan sentido clínico (es decir, para una persona con conocimiento clínico). Este "ojo clínico", sobre el que reflexionaba Federico de Montalvo (de Montalvo Jääskeläinen, 2018) debe exigir recibir información que sea capaz de juzgar. Si en vez de los ejemplos anteriores encontramos que la taquipnea o respiración acelerada es la principal variable predictiva, el modelo pasa a ser una ayuda que da una forma concreta a la experiencia existente sin contradecirla, en la que asumir un cierto grado de generalidad estará dentro de lo razonable y que, desde luego, no realiza inferencias indeseables. Los patrones en los datos son simplemente esto, correlaciones, y es sólo el experto el que puede relacionar esas correlaciones con fenómenos conocidos o con nuevas hipótesis comprobables.

7. Conclusiones: Aprovechemos el potencial de la IA en las ciencias de la salud

La IA tiene un enorme potencial para apoyar las ciencias de la salud, desde el diagnóstico, los modelos predictivos o las aplicaciones a la gestión. Sus principales puntos fuertes, como la crisis del Covid ha manifestado, son la rapidez y el aprovechamiento de los datos.

Sin embargo, los modelos de caja negra tienen una aplicación limitada en las ciencias de la salud, debido a que no son capaces de generar la confianza necesaria para este tipo de aplicaciones. Esta falta de confianza está justificada en los errores (como el sobreajuste y el sesgo algorítmico) que pueden derivarse de problemas en el desarrollo de los modelos o en la preparación de sus datos de entrada.

En algunos casos, no existen alternativas razonables a los modelos de caja negra, como en tratamiento de imagen y voz. Sin embargo, en muchos otros problemas es posible desarrollar alternativas transparentes. En general, resultaría mucho más provechoso intentar centrar los desarrollos de IA en el sector de la salud en la línea de la IA interpretable, que podrían ser aprovechados de manera directa para la confirmación de conocimiento o para la generación de hipótesis nuevas que puedan comprobarse con experimentos posteriores.

Referencias

Ancochea, J., Izquierdo, J. L., Savana COVID-19 Research Group, & Soriano, J. B. (2020). Evidence of gender differences in the diagnosis and management of coronavirus disease 2019 patients: An analysis of electronic health records using natural language processing and machine learning. *Journal of Women's Health.*

Ardakani, A. A., Kanafi, A. R., Acharya, U. R., Khadem, N., & Mohammadi, A. (2020). Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. *Computers in Biology and Medicine,* 103795.

Beck, B. R., Shin, B., Choi, Y., Park, S., & Kang, K. (2020). Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. *Computational and Structural Biotechnology Journal.*

Chalmers, D. J. (2006). Strong and weak emergence. *The Re-Emergence of Emergence,* 244-256.

de Montalvo Jääskeläinen, F. (2018). ¿ Puede la máquina sustituir al hombre?: Una reflexión jurídica sobre el ojo clínico y la responsabilidad en tiempos del big data. *Razón y Fe: Revista Hispanoamericana De Cultura, 278*(1436), 323-334.

Graziani, D., Soriano, J. B., Rio-Bermudez, D., Morena, D., Díaz, T., Castillo, M., et al. (2020). Characteristics and prognosis of COVID-19 in patients with COPD. *Journal of Clinical Medicine, 9*(10), 3259.

Izquierdo, J. L., Ancochea, J., Soriano, J. B., & Savana COVID-19 Research Group. (2020). Clinical characteristics and prognostic factors for intensive care unit admission of patients with COVID-19: Retrospective study using machine learning and natural language processing. *Journal of Medical Internet Research, 22*(10), e21801.

Lalmuanawma, S., Hussain, J., & Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for covid-19 (SARS-CoV-2) pandemic: A review. *Chaos, Solitons & Fractals,* 110059.

Mitchell, M. (2009). *Complexity: A guided tour* Oxford University Press.

Molnar, C. (2020). *Interpretable machine learning* Lulu. com.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206-215.

Zietz, M., & Tatonetti, N. P. (2020). Testing the association between blood type and COVID-19 infection, intubation, and death. *MedRxiv : The Preprint Server for Health Sciences.*