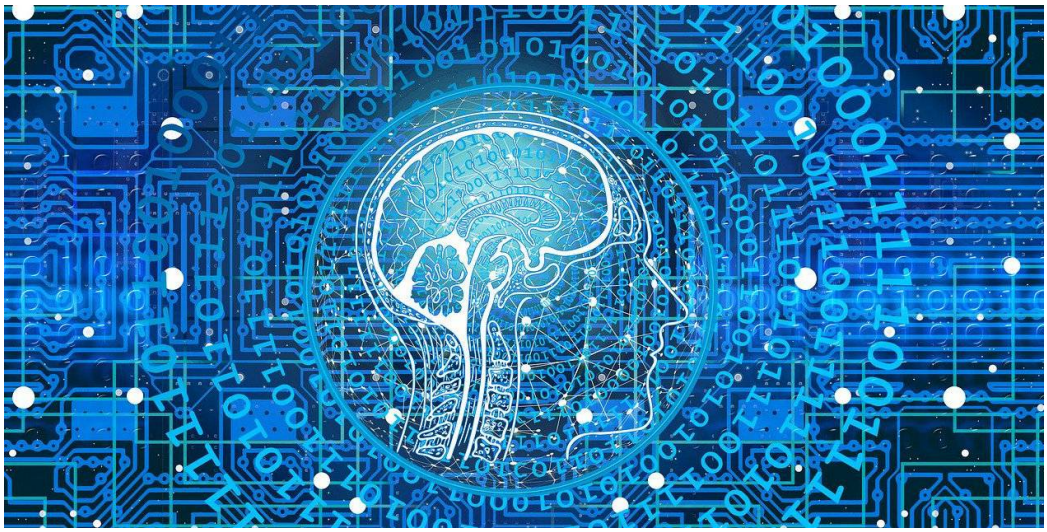
Nuevos algoritmos vigilan el buen comportamiento de la inteligencia artificial

Investigadores de EE UU y Brasil han desarrollado un sistema que ayuda a garantizar que los algoritmos de aprendizaje automático, usados en IA, funcionen adecuadamente y no reproduzcan sesgos discriminatorios. La técnica ha sido probada con éxito en aplicaciones de educación y salud.

SINC 21/11/2019 20:00 CEST



Los investigadores han desarrollado algoritmos que son capaces de entrenar a los algoritmos de aprendizaje automático para que tengan un buen funcionamiento y eviten los sesgos. / Pixabay

La **inteligencia artificial** (IA) tiene cada vez más usos comerciales, gracias a la creciente destreza de los **algoritmos** de aprendizaje automático (***machine learning***, ML), utilizados, por ejemplo, en la conducción de coches autónomos, el control de robots o la automatización de la toma de decisiones.

Pero a medida que la IA se extiende a tareas más delicadas, como puede ser el **diagnóstico médico** o la **selección de personal** para puestos de trabajo, existe una presión creciente para que se ofrezcan **garantías** de que estos sistemas, que se alimentan de **datos históricos**, no reproduzcan **sesgos discriminatorios**.

El sistema permite a los diseñadores de algoritmos de machine learning predecir, con garantías, la fiabilidad de los sistemas alimentados con datos históricos

Ahora, un equipo liderado por las universidades estadounidenses de Stanford y Massachusetts Amherst, en colaboración con la Universidad Federal de Río Grande del Sur, de Brasil, ha desarrollado una técnica que, según sus creadores, permitirá proporcionar estas garantías. El sistema, cuyos resultados se presentan en *Science*, ha sido probado en aplicaciones de IA, en los ámbitos de **educación** y **salud**.

En concreto, los investigadores han desarrollado algoritmos que son capaces de entrenan a los algoritmos de *machine learning* para que tengan un buen funcionamiento y eviten los sesgos.

Emma Brunskill, investigadora de Stanford y autora principal del estudio, indica que con este trabajo pretenden "promover una IA que respete los **valores** de sus usuarios humanos y justifique la confianza que depositamos en los sistemas autónomos".

El estudio se basa en la idea de que si los resultados o comportamientos inseguros o injustos pueden ser definidos matemáticamente, también debería ser posible crear algoritmos que puedan **aprender de los datos** y evitar resultados no deseados con gran fiabilidad.

Predecir la fiabilidad de los algoritmos

Los investigadores también querían desarrollar un conjunto de técnicas que facilitaran a los usuarios de estos algoritmos –que no suelen ser científicos computacionales, sino compañías, centros de investigación, etc.– las especificaciones de los tipos de comportamiento inadecuado que deseen restringir y permitan a los diseñadores de algoritmos de aprendizaje automático predecir, con garantías, la fiabilidad de los sistemas alimentados con datos históricos en IA.

Los algoritmos se han probado con éxito en la mejora de las predicciones de notas de los universitarios para eliminar los sesgos de género

Según comenta a Sinc **Philip Thomas**, científico computacional de la Universidad de Massachusetts Amherst y primer autor del estudio, "con nuestro sistema los diseñadores de algoritmos de aprendizaje automático podrán facilitar a los investigadores, organismos y empresas –que deseen incorporar la IA en sus productos y servicios– la descripción de resultados o comportamientos no deseados que el sistema de IA evitará con alta probabilidad"

Los autores han probado su método en los algoritmos usados en la predicción del promedio de calificaciones de los estudiantes universitarios basada en los resultados de los exámenes, una práctica común que puede resultar en **sesgos de género**. El objetivo era mejorar la imparcialidad de estos algoritmos.

Así, utilizando un conjunto de datos experimentales, dieron **instrucciones matemáticas** a los algoritmos para evitar que realizaran predicciones que sistemáticamente sobreestimaran o subestimaran las calificaciones para un género determinado.

Filtro de imparcialidad

Con estas instrucciones, el algoritmo identificó una mejor manera de predecir las notas de los estudiantes con un sesgo de género mucho menos sistemático que los métodos existentes. Según los investigadores, las técnicas previas mostraron dificultades en este sentido, "debido a que no tenían un **filtro de imparcialidad** incorporado o porque los algoritmos desarrollados para lograr la imparcialidad eran demasiado limitados en su alcance", explica Thomas.

Además, el grupo desarrolló otro algoritmo y lo utilizó para lograr un equilibrio entre seguridad y rendimiento en una **bomba de insulina** automatizada. Estos dispositivos tienen que decidir cuán grande o pequeña

es la **dosis** de insulina que se le debe administrar a un paciente en las comidas.

Lo ideal es que la bomba suministre la insulina suficiente para mantener estables los niveles de azúcar en la sangre. El *machine learnig* puede ayudar a identificar **patrones** sutiles en las respuestas del **azúcar en sangre** de una persona a las dosis, pero los sistemas existentes no facilitan a los médicos la especificación de los resultados que los algoritmos de dosificación automatizada deben evitar, como los colapsos por bajo nivel de azúcar en la sangre.

Otro algoritmo fue aplicado en la mejora del funcionamiento de la dosificación con IA de una bomba de insulina

Utilizando un simulador de glucosa en sangre, Brunskill y Thomas mostraron cómo se podían entrenar las bombas para identificar la dosificación adaptada para esa persona, evitando complicaciones por sobredosis o subdosis. Aunque el grupo no está listo para probar este algoritmo en humanos, apunta a un enfoque de IA que podría mejorar la calidad de vida de los diabéticos, dicen los investigadores.

Método seldoniano

En su artículo de *Science*, los autores usan el término 'algoritmo seldoniano' para definir su enfoque. Hacen referencia a **Hari Seldon**, un personaje inventado por **Isaac Asimov**, que una vez proclamó las tres **leyes de la robótica**, comenzando con el mandato de que "un robot no puede herir a un ser humano o, a través de la inacción, permitir que un ser humano resulte perjudicado".

Thomas señala que, aunque aún queda mucho por hacer, el nuevo **método seldoniano** "facilitará a los diseñadores de algoritmos de aprendizaje automático la creación de instrucciones para evitar el comportamiento inadecuado de estos algoritmos, de manera que les permita evaluar la probabilidad de que los sistemas entrenados funcionen correctamente en el

mundo real.

Este trabajo ha recibido financiación parcial de Adobe, la Fundación Nacional de Ciencias y el Instituto de Ciencias de la Educación de EE UU.

Referencia bibliográfica:

Philip S. Thomas, Bruno Castro da Silva, Andrew G. Barto, Stephen Giguere, Yuriy Brun, Emma Brunskill."Preventing undesirable behavior of intelligent machines". *Science* (21 noviembre, 2019)

TAGS | INTELIGENCIA ARTIFICIAL | MACHINE LEARNING | ALGORITMOS | SESGOS |

La ciencia es noticia