Generación sintética de datos clínicos

1. El uso de datos sintéticos en salud

El campo biomédico es uno de los sectores más afectados por la creciente regulación de la Inteligencia Artificial (IA) y la legislación sobre protección de datos, dada la sensibilidad de la información de los pacientes. Varios gobiernos están introduciendo normativas estrictas para el tratamiento de datos personales y las aplicaciones de IA, como la nueva ley de IA de la Unión Europea, la CCPA (Estados Unidos) y la LGPD (Brasil). En el ámbito de la investigación biomédica, es preciso actuar con cautela a la hora de emplear datos de pacientes para el entrenamiento de modelos de IA. Los datos de pacientes, caracterizados por su naturaleza sensible, están sujetos a una estricta protección que exige preservar la privacidad. Una potencial solución que puede superar estas limitaciones es la generación de datos totalmente sintéticos como alternativa a los datos reales.

Los datos sintéticos son datos artificiales generados por un modelo entrenado y construido para replicar los datos reales teniendo en cuenta su distribución (media, varianza) y estructura (por ejemplo, la correlación entre atributos). Este nuevo paradigma emerge como una metodología versátil en el aprendizaje automático, extendiendo sus aplicaciones a dos dominios: salvaguardar la privacidad de la información sensible y aumentar los conjuntos de datos para mejorar el entrenamiento de modelos.

En los métodos de generación de datos sintéticos, existe un equilibrio delicado entre el realismo de los datos generados y la protección de la privacidad de los individuos. Por un lado, para que los datos sintéticos sean útiles y efectivos en aplicaciones médicas, es crucial que reflejen con precisión las características y distribuciones de los datos reales. Esto significa que los datos sintéticos deben ser lo suficientemente realistas como para que los modelos de Inteligencia Artificial puedan funcionar correctamente en situaciones del mundo real. Por otro lado, la privacidad de los individuos debe ser protegida de manera robusta. Esto implica que los datos sintéticos no deben contener información identificable que pueda vincularse con individuos reales. El objetivo es preservar la confidencialidad y la seguridad de los datos mientras se garantiza su utilidad para construir modelos de IA. Este equilibrio entre realismo y privacidad puede ser difícil de lograr. A veces, aumentar la precisión de los datos sintéticos puede implicar un mayor riesgo de identificación de individuos. Por lo tanto, es fundamental encontrar métodos de síntesis que logren un alto nivel de realismo sin comprometer la privacidad.

Otra de las aplicaciones clave de los datos sintéticos es la ampliación de los conjuntos de datos existentes, lo cual mejora significativamente el entrenamiento de los modelos de IA. La construcción de modelos de IA robustos y precisos requiere grandes cantidades de datos diversos, una condición que a menudo no se cumple en las bases de datos biomédicas. Estas bases de datos suelen contener un número limitado de muestras, especialmente en enfermedades raras, donde la recolección de datos es especialmente difícil por la baja prevalencia de estas condiciones. Los datos sintéticos pueden replicar las características estadísticas y clínicas de los datos reales, permitiendo así la generación de grandes volúmenes de datos que de otro modo no estarían disponibles. Esto no solo mejora la

capacidad de los modelos para aprender y generalizar a partir de la información proporcionada, sino que también ayuda a abordar problemas de sesgo y sobreajuste que pueden surgir cuando se trabaja con conjuntos de datos pequeños.

Por estos motivos, los datos sintéticos emergen como un nuevo paradigma para superar las limitaciones actuales de la IA, facilitando avances significativos en el ámbito de la medicina. Esta innovadora tecnología no solo permite el acceso a conjuntos de datos más amplios y diversos, sino que también garantiza la privacidad y seguridad de la información sensible. Así, los datos sintéticos allanan el camino hacia un progreso médico más rápido y eficiente, abriendo nuevas oportunidades para la investigación y el desarrollo de tratamientos.

2. Tipos de datos sintéticos

Los datos sintéticos pueden clasificarse en dos categorías principales: estructurados y no estructurados.
Los datos sintéticos estructurados se refieren a datos organizados en un formato tabular, similar a los datos que se encuentran en bases de datos relacionales. Estos datos se componen de filas y columnas, donde cada columna representa una variable específica y cada fila corresponde a una instancia de datos.
La generación de datos sintéticos tabulares puede realizarse mediante diversas técnicas:

1. **Modelos Probabilísticos:** Emplean distribuciones estadísticas para generar datos. Por ejemplo, se puede utilizar una distribución normal para generar una variable continua o una distribución binomial para variables categóricas.
2. **Modelos de Machine Learning:** Algoritmos avanzados como Redes Generativas Adversariales (GANs) tabulares pueden ser entrenados para generar datos que imiten las características de los datos reales.

Los datos sintéticos no estructurados se refieren a datos que no siguen un formato fijo o tabular, tales como imágenes, texto, audio y video. Estos datos suelen ser más complejos de generar debido a su naturaleza altamente dimensional y variada.
Las principales técnicas para generar imágenes son:

1. **Redes Generativas Adversariales (GANs):** Estas redes utilizan una arquitectura compuesta por dos redes (generadora y discriminadora) que compiten entre sí para crear imágenes realistas. Ejemplos incluyen StyleGAN, que puede generar rostros humanos que son difíciles de distinguir de los reales.
2. **Modelos de Difusión:** Estos modelos generan imágenes mediante un proceso iterativo de refinamiento, donde se parte de ruido aleatorio y se va ajustando progresivamente hasta obtener una imagen coherente. Ejemplos notables incluyen DALL-E 2 y Stable Diffusion, que son capaces de crear imágenes a partir de descripciones textuales.

Las principales técnicas para generar texto se basan en:

1. **Modelos de Lenguaje:** Algoritmos como GPT-4 pueden generar texto coherente y relevante entrenándose en grandes volúmenes de datos textuales reales. Estos modelos aprenden patrones lingüísticos y contextuales para producir nuevo texto.

Existen muchos métodos para generar datos sintéticos, cada uno con sus propias ventajas y aplicaciones únicas. La precisión y realismo de los datos sintéticos pueden variar, y la generación de datos verdaderamente indistinguibles de los reales sigue siendo un desafío.

3. Open Data

El concepto Open Data, o Datos Abiertos, hace referencia a una filosofía de acumulación de datos electrónicos que persigue su divulgación, con pocas o ninguna restricción y permitiendo su uso por parte de terceros, con el simple objetivo de ayudar a cualquier persona, entidad u organización que los necesite para desarrollar una actividad, un proyecto de investigación, un negocio o con cualquier otra finalidad.

El movimiento de Datos Abiertos se consolidó cuando en diciembre de 2007 una treintena de expertos del gobierno abierto de los Estados Unidos estableció “los 8 principios del Open Data”, que se resumen en lo siguiente:

1. **Libertad.** Los datos abiertos serán públicos y accesibles, y bajo ningún concepto se restringirá su acceso por privacidad, seguridad, privilegios o derechos de autor.
2. **Originalidad.** Los datos abiertos serán presentados como originalmente se encontraban cuando se recolectaron, lo más detallados posible y sin haber sido modificados ni transformados. En definitiva, los datos abiertos son datos en bruto.
3. **Actualidad.** Los datos abiertos serán presentados lo antes posible, de forma que sean útiles en el momento de su publicación y no “caduquen” por el paso del tiempo.
4. **Accesibilidad.** Los datos abiertos estarán disponibles al mayor número de personas para el mayor número de propósitos. Para su publicación se debe de utilizar Internet y permitir la descarga de estos de manera sencilla, para que todos podamos hacerlo.
5. **Claridad.** Los datos abiertos tendrán una estructura que permitirá a programas informáticos su procesamiento automático. Esto significa que los datos estarán en formato texto limpio (“Bloc de notas” de Windows) y con suficientes explicaciones de qué significa cada característica de los datos recogidos.
6. **Libre descarga.** Los datos abiertos se podrán descargar sin registrarse en ningún formulario ni dejar constancia de quién ha accedido a ellos. Es decir, se garantizará el acceso anónimo a los datos.
7. **Dominio público.** Los datos abiertos no serán propiedad de nadie; serán propiedad de todos. De esta manera, ninguna persona ni organismo podrá restringir quién puede ver y utilizar los datos.
8. **Sin licencia.** Los datos abiertos no pueden ser licenciados por copyright, patentes o marcas comerciales. Pueden someterse a ciertas medidas de seguridad y privacidad, pero nunca pueden ser vendidos o intercambiados “en secreto”.

En resumen, los datos abiertos son un material que los proyectos de investigación, instituciones gubernamentales y negocios pueden utilizar libremente y sin ninguna restricción, salvo la de mantener y perpetuar el “trato” o “acuerdo” de los datos abiertos (es decir, que sigan estando disponibles para todos y los demás principios descritos anteriormente).

En este programa, vosotros, los estudiantes, vais a beneficiaros de esta filosofía de los datos abiertos para realizar un pequeño proyecto de investigación en el área sanitaria. Todo proyecto de investigación, especialmente si utiliza Inteligencia Artificial, necesita de muchos datos; y en esto está incluido también el generar nuevos datos, como se ha discutido con anterioridad. Es por ello por lo que cada vez más gobiernos europeos y empresas están impulsando este movimiento de compartir datos, pues con ello estamos también compartiendo conocimiento en beneficio de todos.

4. Fuentes de datos abiertos

La búsqueda de datos abiertos no es una tarea compleja en un principio, pues podemos rápidamente llegar a muchas fuentes de datos con una búsqueda en Interenet. Sin embargo, debemos asegurarnos de que efectivamente los datos que estamos viendo pueden ser usados para la finalidad que nos interesa. Para identificar qué datos podemos usar, debemos fijarnos en en su Licencia y en sus Condiciones. Podemos encontrar fuentes de datos abiertos si nos fijamos en que den su consentimiento expreso a su libre uso o en que estén licenciadas bajo aquellas licencias “Creative Commons” que no prohíban su distribución y uso.

En nuestro caso, estamos buscando fuentes de datos que contengan imágenes, textos o datos tabulares. Algunas de las fuentes más conocidas de imágenes [4] (también llamados “repositorios”) son:

- MNIST, que contiene más de 70.000 imágenes en blanco y negro de números escritos a mano para entrenar algoritmos de reconocimiento de números.
- ImageNET, que recoge palabras o conjuntos de palabras con un significado asociado a cientos o miles de imágenes.

Sin embargo, existen cientos de fuentes de las que podemos extraer imágenes que nos serán útiles. Lo más importante es asegurares de que su uso esté permitido y que efectivamente se trata de imágenes, texto o datos tabulares que podemos usar.

5. Preguntas fundamentales

Como primer objetivo de proyecto, queremos contestarnos las siguientes preguntas:

- ¿Qué problemas podría resolver el uso de datos sintéticos médicos en la investigación?
- ¿Qué tipo de datos, desde el punto de vista médico, se podrían generar, de forma que resultaran útiles?
- ¿En qué formato y por qué canales se distribuyen datos biomédicos abiertos?
- ¿De qué forma se suelen obtener y extraer los datos biomédicos de los bancos de datos?
- ¿Consideráis ético el almacenamiento de datos biomédicos de los pacientes de un hospital?
- **¿Estáis conformes con el uso de Inteligencia Artificial en los hospitales?**

6. Trabajo y entrega

A la hora de abordar este proyecto de investigación, queremos que tengáis la libertad de explorar y buscar información por vuestra cuenta de forma que gracias a lo que habéis aprendido, elaboréis un trabajo que explore aquellas secciones de la generación de datos sintéticos clínicos que os parezcan más interesantes u os hayan llamado más la atención. El objetivo no es solo que vayáis más allá e investiguéis más acerca del tema por vuestra cuenta, sino que además tengáis la posibilidad de demostrar todo lo aprendido de la forma que veáis que se adapte más al tema y a vosotros mismos. Podéis usar las preguntas del punto anterior como puntos de partida que os sirvan de ayuda a la hora de empezar, pero valoraremos que seáis originales y demostréis vuestra creatividad.

La extensión de los trabajos es libre, pero a modo orientativo, se espera un **mínimo de 12 páginas y un máximo de 25,** portada, índice y página de referencias **incluidas**.

A continuación, os dejamos un ejemplo de una posible estructura del trabajo para que os sirve de posible guía:

- Definir **dato sintético** y descubrir de qué tipos existen, para qué se utilizan tradicionalmente en el mundo de la medicina y qué interés tienen en el mundo de la IA.
- Investigar y especular un poco sobre el impacto que pueden llegar a tener la generación de datos sintéticos en los hospitales y equipos de investigación de medicina. No hace falta dar detalles concretos; sólo divagar un poco en qué se podría lograr con ello.
- Estudiar las principales dificultades a la hora de generar datos sintéticos para los hospitales, tanto si son tecnológicas, como éticas o legales.
- Investigar y nombrar modelos de IA se pueden usar para generar datos sintéticos, y explicar sus conceptos más básicos.
- Responder a las preguntas fundamentales listadas en la sección 5.

Referencias:

[1] What is Computer Vision? | IBM. https://www.ibm.com/es-es/topics/computer-vision

[2] Saiz, F. A., & Barandiarán, Í. (2020). COVID-19 Detection in Chest X-ray Images using a Deep Learning Approach. International Journal of Interactive Multimedia and Artificial Intelligence, 6(2), 4. https://doi.org/10.9781/ijimai.2020.04.003

[3] Cohen, J. P., Morrison, P., & Dao, L. (2020). COVID-19 image data collection. arXiv preprint arXiv:2003.11597. https://doi.org/10.48550/arXiv.2003.11597

[4] datos.gob.es. (2022, October 19). Repositorios abiertos de imágenes para entrenamiento de modelos de Inteligencia Artificial. datos.gob.es. https://datos.gob.es/es/blog/repositorios-abiertos-de-imagenes-para-entrenamiento-de-modelos-de-inteligencia-artificial

[5] Macias-Fassio, E., Morales, A., Pruenza, C., & Fierrez, J. (2024). Privacy-Preserving Statistical Data Generation: Application to Sepsis Detection. *arXiv preprint arXiv:2404.16638*. https://arxiv.org/abs/2404.16638

[6] Murtaza, H., Ahmed, M., Khan, N. F., Murtaza, G., Zafar, S., & Bano, A. (2023). Synthetic data generation: State of the art in health care domain. *Computer Science Review*, *48*, 100546. https://www.sciencedirect.com/science/article/pii/S1574013723000138