Diseño de un modelo generador de datos sintéticos para el análisis de big data de investigación médica.

Alfonso Esteban Lasso[a], Cristina Martínez Toledo, Sergio Perosanz Amarillo

Inari Biotech, Dpto. Biomedicina, Universidad de Alcalá. Universidad Politécnica de Madrid.

a. inaribiotech@gmail.com

VIII Congreso de Señalización Celular, SECUAH 2022.
21 a 24 de marzo, 2023. Universidad de Alcalá. Alcalá de Henares, Madrid. España.

Palabras clave: Bioinformática; inteligencia artificial; datos sintéticos; ensayos clínicos

Resumen

La escasez de datos o una distribución irregular, con espacios poco poblados de datos, son dificultades frecuentes de la investigación biomédica. Esta escasez de información dificulta el trabajo estadístico, impidiendo la obtención de conclusiones válidas. Las enfermedades raras, que se definen como aquellas enfermedades cuya prevalencia es inferior a 5 por cada 10.000 personas, son un ejemplo paradigmático de escasez de datos debido a su baja prevalencia. Actualmente es posible solucionar parcialmente este problema mediante la utilización de datos sintéticos. Los datos sintéticos son datos fiables generados mediante inteligencia artificial que complementan a los datos reales cuando los conjuntos de datos reales carecen de calidad, volumen o variedad. En este estudio se compararon la calidad de los datos sintéticos de investigación médica obtenidos mediante distintos algoritmos generadores de datos sintéticos. Se utilizaron datos tabulares de ensayos clínicos oncológicos procedentes del banco de datos de Project Data Sphere (PDS) y del National Cancer Institute. Algunos de los indicadores usados para comparar los datos sintéticos con los reales y evaluar la calidad de los datos sintéticos generados fueron: precisión, recall y F1score. Según nuestro estudio, la puntuación de los algoritmos menos intensivos en computación era relativamente baja; 63% para GaussianCopula, y 68% para Fast-ML. Los algoritmos más sofisticados generaron una mejor puntuación promedio; 74% para CTGAN, 78% para CopulaGAN y 82% para TVAE. En conclusión, los algoritmos TVAE son los más idóneos para generar datos sintéticos de datos médicos y serian de utilidad, por ejemplo, para generar datos médicos de pacientes con enfermedades raras que aumenten la base de estudio. Además, se desarrolló un nuevo algoritmo sobre la base de TVAE para su uso en investigación médica y en ensayos clínicos. El algoritmo se probó para generar datos sintéticos con los datos del ensayo N0147 de PDS. Este trabajo demuestra que se obtienen las mismas conclusiones analizando los datos sintéticos que analizando los datos reales lo cual sugiere que los datos sintéticos podrían ser de utilidad en investigación médica. Inari agradece a la CAM la concesión de ayudas del programa Investigo.

Cita: Esteban Lasso, Alfonso; Martínez Toledo, Cristina; Perosanz Amarillo, Sergio (2023) Diseño de un modelo generador de datos sintéticos para el análisis de big data de investigación médica. Actas del VIII Congreso de Señalización Celular, SECUAH 2022. 21 a 24 de marzo, 2023. Universidad de Alcalá. Alcalá de Henares, Madrid. España. dianas 12 (1): e202303ev01. ISSN 1886-8746 (electronic)
journal.dianas.e202303ev01 https://dianas.web.uah.es/journal/e202303ev01.
URI http://hdl.handle.net/10017/15181