Análisis de sesgos en algoritmos de clasificación de oximetría de pulso para detección de trastornos del sueño

Juan Manuel Perero[1], Enzo Ferrante[1], Luis Larrateguy[2], Leandro Di Persia[1], and Hugo Leonardo Rufiner[1]

[1] Instituto de investigación en señales, sistemas e inteligencia computacional (sinc(i)), FICH-UNL/CONICET, Ciudad Universitaria UNL, (S3000) Santa Fe, Argentina
[2] Centro de Medicina Respiratoria de Paraná, Entre Ríos, Argentina
jmperero@sinc.unl.edu.ar
www.sinc.unl.edu.ar/

Abstract. La utilización de algoritmos y técnicas de aprendizaje automático dentro del ámbito de la medicina ha demostrado ser de gran utilidad en tareas de asistencia al diagnóstico. Al mismo tiempo, durante los últimos años, se han dado a conocer problemáticas importantes que afectan el funcionamiento de estos métodos en subpoblaciones específicas, presentando un rendimiento dispar en determinados grupos demográficos. Este rendimiento dispar suele estar asociado a la subrepresentación de dichas poblaciones en los datos utilizados durante el entrenamiento, resultando en modelos sesgados. Este es un trabajo preliminar que busca explorar la existencia de sesgos en la clasificación de apnea a partir de oximetría de pulso considerando diferentes grupos étnicos. Para ello, se utiliza una base de datos con información étnica de pacientes y un algoritmo para detectar trastornos de sueño como apneas o hipopneas. Para la detección de dichos eventos, se emplea sólo la señal de saturación periférica de oxígeno en sangre (SpO2) junto al algoritmo basado en aprendizaje de diccionarios DAS-KSVD. Los experimentos consisten en analizar la base de datos de polisomnografía MESA, que contiene cuatro grupos étnicos, utilizando la señal de SpO2 para el aprendizaje de diccionarios con los que se mejora el proceso de detección de patologías.

Keywords: Aprendizaje de diccionarios · detección de trastornos del sueño · sesgo.

1 Introducción

El uso de algoritmos computacionales en el diagnóstico médico ha crecido con los avances en inteligencia artificial (IA). Sin embargo, se ha advertido que estos modelos pueden tener sesgos en subpoblaciones, lo cual es especialmente preocupante en datos de salud. [2]. Como consecuencia de dichos estudios, el campo de la equidad algorítmica (o *fairness* en inglés) ha comenzado a ganar mayor importancia en los últimos años, particularmente en el caso de la medicina [5].

Una de las contribuciones más importantes de los estudios presentados en este área es comprender el comportamiento y rendimiento de los modelos de aprendizaje automático frente a distintos grupos poblacionales, atributos protegidos o características sensibles. Recientemente, se ha reportado que los oxímetros de pulso, sensores con los que se mide el oxigeno en sangre, presentan variaciones relacionadas a la tonalidad de la piel de los pacientes [3]. Dichas variaciones son resultado del empleo de luz por parte de los sensores y de distintos niveles de melanina en la piel de los diversos grupos. En este trabajo se pretende analizar el impacto de esta variabilidad, a través de experimentos para evaluar la existencia de sesgos al entrenar un clasificador sobre datos en diferentes etnias. El estudio a realizar se centra en el uso de la *saturación periférica de oxigeno* (SpO2) obtenida por oximetría de pulso, para la detección de trastornos de sueño como apneas e hipopneas. El *Síndrome de Apnea/Hipopnea obstructiva* (OSAH por sus siglas en inglés), implica una obstrucción de las vías respiratorias de forma total o parcial durante el ciclo de sueño. Para su diagnóstico, el estudio de preferencia a realizarse es la polisomnografía que implica un análisis de la etapa del sueño realizando múltiples de mediciones, como electrocardiograma, SpO2, flujo de aire, entre otros. Siendo la polisomnografía un estudio complejo y costoso, en [8] se plantea reducir la cantidad de señales empleando solo SpO2. Una técnica destacada en esta tarea, enmarcada en el área aprendizaje de diccionarios, es el algoritmo DAS-KSVD [6], el cual será evaluado en este trabajo.

2 Materiales y Métodos

En la presente sección, se brindará una presentación de los datos y los métodos empleados en la ejecución de la investigación.

Base de datos MESA. Multi-Ethnic Study of Atherosclerosis (MESA) [1] es la base de datos utilizada y cuenta con polisomnografias de 2056 pacientes, realizadas a hombres y mujeres de entre 45 a 84 años. Entre el conjunto de metadatos que se compila para cada paciente, se encuentra información demográfica referida al grupo étnico, que será útil a los fines de este trabajo. La base de datos cuenta con 4 grupos étnicos (830 Caucásicos, 616 Afroamericano, 526 Latinoamericano, 265 Chinoamericano) que serán considerados como atributos protegidos.
Algoritmo DAS-KSVD. "Discriminant Atom Selection KSVD" (DAS-KSVD) [6] es un método que utiliza aprendizaje de diccionario, seleccionando átomos discriminantes para cada clase del problema. Estos átomos se extraen de diccionarios generados por el algoritmo *"K Singular Value Decomposition"*(KSVD) y se evalúa su grado de discriminabilidad basado en la frecuencia y magnitud de activación para la partición de entrenamiento. El método construye iterativamente diccionarios para cada clase a representar, y luego usa las activaciones del diccionario conjunto final como características de entrada a un perceptrón multicapa (MLP por sus siglas en inglés). Dicho MLP será entrenado para predecir la clase a la que pertenece un segmento, siendo en este caso las posibles clases: normal o con eventos de apnea/hipoapnea.

2.1 Análisis de sesgo en clasificación de señales de oximetría

Pre-procesamiento y particionado de los datos: Como la señal SpO2 extraída de la base de datos posee larga duración y errores por desconexión, lo primero a realizarse fue la corrección de errores, filtrado y segmentado (de forma similar a [7]). Posteriormente, se realizó un particionado de los conjuntos en entrenamiento, validación y prueba para cada una de las etnias de trabajo. Se tomó un 70% de los pacientes disponibles para entrenamiento, 10% para validación y 20% para test. Para la realización de múltiples experimentos se utilizó validación cruzada K-fold con un K=10.
Entrenamiento de los modelos: Este trabajo preliminar se enfocó en el análisis de la existencia de sesgo considerando sólo dos etnias, caucásica y afroamericana. Para analizar los escenarios contrafácticos, se entrenaron 3 modelos, uno sólo con datos de pacientes caucásicos, otro sólo con afroamericanos y el tercero mixto, donde se mezclan los dos grupos a partes iguales. En la etapa de entrenamiento, se utilizó el conjunto de datos para generar un diccionario con el algoritmo DAS-KSVD. Una vez aprendido, los mismos datos se proyectaron sobre este diccionario y se entrenó un MLP para realizar la clasificación.
Evaluación de sesgo: Para analizar la existencia de sesgo, se siguió una metodología de análisis contrafáctico similar a la propuesta en [4]. Dado un conjunto de prueba de una etnia específica, la estrategia consiste en comparar el rendimiento de los 3 modelos entrenados anteriormente. La hipótesis inicial es que, si la tarea bajo análisis es propensa a sesgarse respecto al atributo protegido *etnia*, entonces el mejor desempeño debería darse con el clasificador entrenado con la misma etnia con la que se evalúa, y la peor con el modelo correspondiente a una etnia diferente. Las métricas utilizadas en la evaluación son F1-score, el área bajo la curva característica operativa del receptor (AUC-ROC) y el área bajo la curva "Precision-Recall" (AUC-PR), contemplando el desbalance en las clases normal-evento de 60-40%.

3 Resultados y discusión

Contrario a lo esperado, el análisis no reveló de manera significativa la presencia de sesgos en las diversas combinaciones de grupos. Si bien en este trabajo se presentan resultados para el caso Caucásico-Afroamericano, el resto de las combinaciones presentan la misma tendencia. En primera instancia se analiza el promedio de las matrices de confusión (ver Figura 1) a partir de las cuales no es posible establecer una tendencia clara. Adicionalmente, se analizaron los valores de F1-score, AUC-ROC y AUC-PR utilizando la misma metodología. En este caso, la observación de los conjuntos evaluados, no parecen existir diferencias significativas al variar los conjuntos de entrenamiento. Basándose en los experimentos preliminares presentados, no fue posible establecer la existencia sistemática de sesgo para la base de datos MESA respecto al grupo étnico para el problema de detección de trastornos del sueño con el método DAS-KSVD. Sin embargo, esto no significa que los mismos no existan. A futuro, se pretenden explorar otras bases de datos y otro tipo de estrategias para la clasificación (como

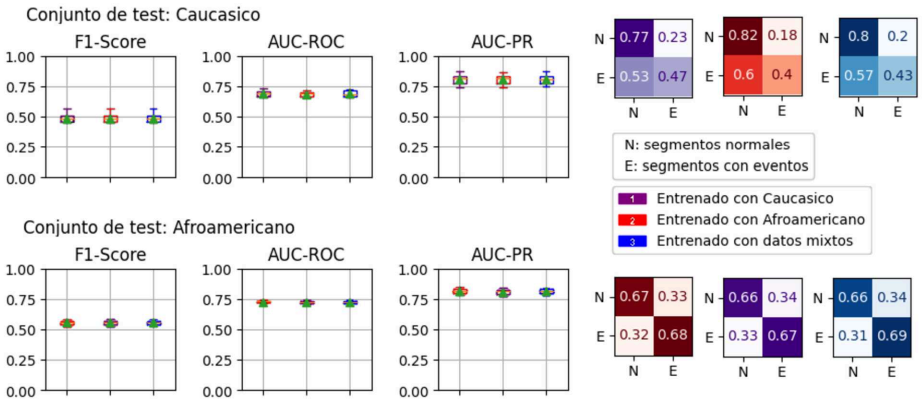

Fig. 1. Resultados del análisis contrafáctico, los resultados preliminares no arrojan una tendencia clara de potencial existencia de sesgos.

redes neuronales convolucionales) para entender si existen modelos más propensos que otros a sesgarse. Otra cuestión a explorar es el balance de los datos a nivel de clases entre los segmentos de K-fold implementados, para garantizar que el único factor de variación en los datos de entrenamiento sea el grupo étnico.

Referencias

1. Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R., Diez Roux, A.V., Folsom, A.R., Greenland, P., JacobsJr, D.R., Kronmal, R., Liu, K., et al.: Multi-ethnic study of atherosclerosis: objectives and design. American Journal of epidemiology **156**(9), 871–881 (2002)
2. Chen, I.Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., Ghassemi, M.: Ethical machine learning in healthcare. Annual Review of Biomedical Data Science **4**(1), 123–144 (2021). https://doi.org/10.1146/annurev-biodatasci-092820-114757
3. Keller, M.D., Harrison-Smith, B., Patil, C., Arefin, M.S.: Skin colour affects the accuracy of medical oxygen sensors (Oct 2022). https://doi.org/10.1038/d41586-022-03161-1, https://www.nature.com/articles/d41586-022-03161-1
4. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc. of the Nat. Acad. of Sciences **117**(23), 12592–12594 (2020)
5. Ricci Lara, M.A., Echeveste, R., Ferrante, E.: Addressing fairness in artificial intelligence for medical imaging. Nature Communications **13**(1), 4581 (2022)
6. Rolon, R.E., Di Persia, L.E., Spies, R.D., Rufiner, H.L.: A multi-class structured dictionary learning method using discriminant atom selection. Pattern Analysis and Applications **24**(2), 685–700 (2021)
7. Rolón, R.E., Gareis, I.E., Larrateguy, L.D., Di Persia, L.E., Spies, R.D., Rufiner, H.L.: Automatic scoring of apnea and hypopnea events using blood oxygen saturation signals. Biomedical Signal Processing and Control **62**, 102062 (2020)
8. Terrill, P.I.: A review of approaches for analysing obstructive sleep apnoea-related patterns in pulse oximetry data. Respirology **25**(5), 475–485 (2020)