Sesgos en problemas de regresión originados por el desbalance de datos en términos de atributos protegidos

Estanislao Claucich[1], Enzo Ferrante[1]*, Rodrigo Echeveste[1]*

[1] Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), FICH-UNL/CONICET, Argentina

Resumen En este trabajo se busca estudiar el efecto sobre el desempeño de modelos de regresión provocado por el desbalance en los datos en términos de atributos protegidos durante el entrenamiento. Estos atributos, como género o color de piel de una persona, son características propias de los datos que pueden o no tener una relación directa con el problema a resolver. Los resultados obtenidos mediante experimentos tanto sobre datos sintéticos como reales, muestran que el error sobre una dada población aumenta cuando se encuentra subrepresentada en el conjunto de datos de entrenamiento. En ambos casos estudiados encontramos que el error sobre la población completa fue mínimo cuando se encontraba balanceado en términos del atributo protegido en cuestión. Este estudio es el primer paso de un trabajo que busca extender este análisis a otras bases de datos, modelos y problemas, para luego atenuar este inconveniente incorporando penalizadores que desincentiven un mejor rendimiento sobre un subconjunto en desmedro de otro.

Palabras claves: justicia algorítmica, aprendizaje profundo, sesgos

1. Introducción

El creciente desarrollo de algoritmos de inteligencia artificial (IA) hizo que los mismos se encuentren cada vez más presentes en diversas disciplinas y aplicaciones, incrementando el potencial impacto social de estas tecnologías. De esta forma se vuelve cada vez más importante analizar posibles problemas, riesgos o incluso fallas que estos algoritmos puedan generar a la hora de ser utilizados. Una particularidad es que su funcionamiento estará en parte determinado por los datos que se utilizan para su confección, por lo que es de suma importancia analizar el efecto que estos tienen sobre el desempeño de los modelos, ya que podrían contener sesgos que luego sean transferidos a los sistemas de IA [10,2].

De esta forma, el concepto de *justicia algorítmica* en IA es de gran interés para distintos grupos de investigación, los cuales se centran en estudiar diferencias en el desempeño de los algoritmos de IA en función de atributos protegidos y diseñar soluciones que permitan reforzar la confianza de la sociedad en estos

∗ Iguales Contribuciones

sistemas [4]. Distintos estudios demuestran que el uso de bases de datos desbalanceadas en cuanto a determinados atributos protegidos puede impactar directamente en el rendimiento de los modelos, presentando un menor rendimiento en los grupos subrepresentados [5,3].

Este trabajo extiende lo realizado en [1] donde se analizó el impacto del desbalance de género en un problema de *clasificación* con *datos sintéticos*. En este caso se busca caracterizar este efecto sobre problemas de *regresión*, tanto para datos sintéticos como para *datos reales*.

2. Resultados

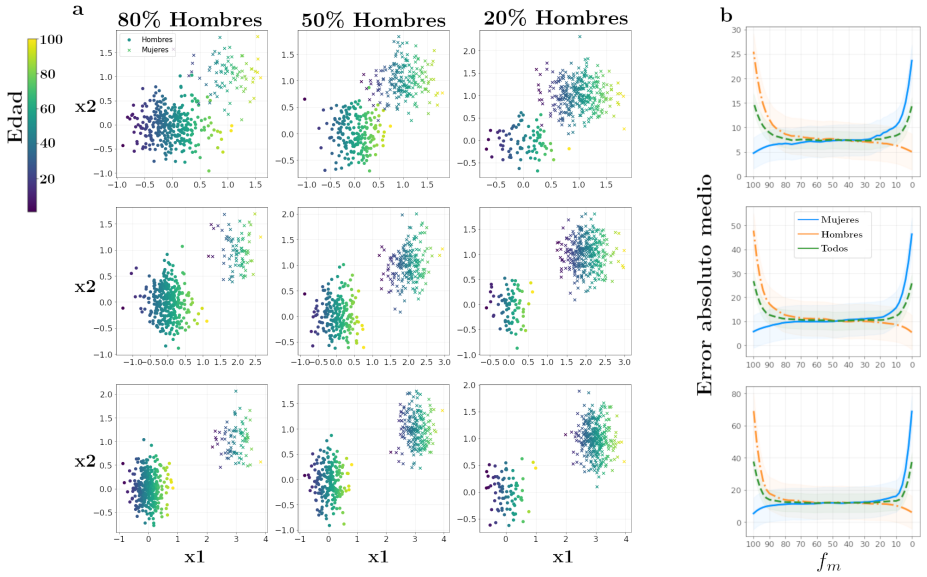

Figura 1: **Error Absoluto Medio (EAM) de un regresor en función del desbalance de género en el conjunto de entrenamiento**
a, Datos de entrenamiento de hombres (círculos) y mujeres (cruces), en tres casos distintos de tensión entre los conjuntos (filas). $x1$ y $x2$ corresponden a características ficticias que representan a la población sintética bajo análisis. El color de cada punto representa la edad objetivo de un individuo. **b**, EAM del regresor lineal ($\pm$ un desvío estándar) en el conjunto de prueba en función de la proporción de mujeres durante el entrenamiento (f_m). Se muestran los resultados para cada subconjunto (mujeres en azul y hombres en naranja) y para la población completa (en verde), para los distintos casos de tensión.

Primeramente se entrenó un regresor lineal usando datos sintéticos de hombres y mujeres, generados de modo que el regresor ideal correspondiente a cada

sub-grupo no sea idéntico debido a un corrimiento en la variable a inferir. Se estudiaron diferentes configuraciones de distribuciones de entrenamiento desbalanceadas (Fig. 1 a), calculando luego el error absoluto medio (EAM) sobre un conjunto de prueba balanceado con respecto al atributo protegido. Se observa cómo el modelo se ajusta de mejor manera para aquel subconjunto que se encuentra mayormente representado durante el entrenamiento, deteriorándose el rendimiento para la otra población (Fig. 1 b). A su vez, a mayor tensión entre estos subconjuntos mayor es el sesgo entre los subgrupos. Es importante notar que el desbalance del atributo protegido disminuye el rendimiento sobre la población total, siendo mínimo el EAM cuando los subconjuntos se encuentran balanceados.

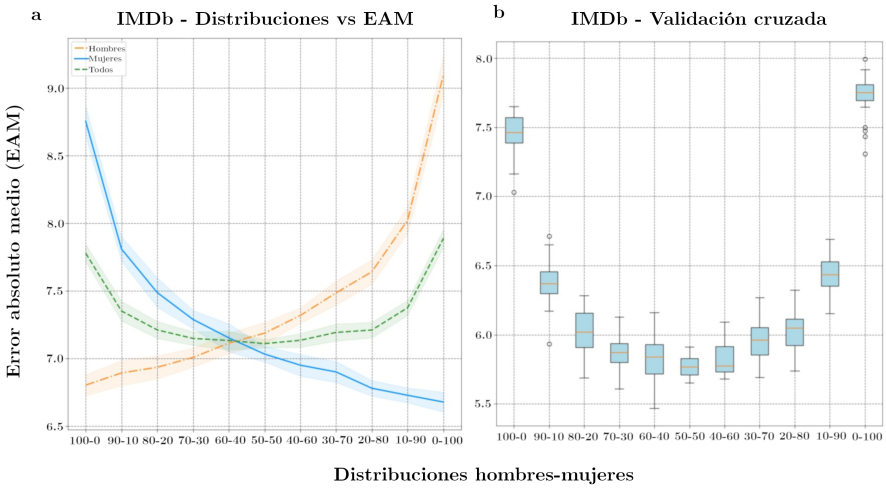

Figura 2: **EAM de IMDb en función del desbalance de género durante el entrenamiento.**
a, EAM (± un desvío estándar) de las distintas poblaciones individuales (naranja para hombres y azul para mujeres) y para la población completa (en verde). **b**, Validación cruzada para cada una de las distribuciones analizadas.

Luego se estudió la tarea de predecir la edad a partir de una foto del rostro de una persona, usando actores y actrices de diferentes edades de la base de datos de IMDb [6]. Se realizó un ajuste fino de las últimas 4 capas de una red neuronal VGG-19 pre-entrenada con ImageNet [7], entrenando un ensamble de 5 modelos utilizando, para cada uno, distintos grados de desbalance entre hombres y mujeres (Fig. 2). Se llevó a cabo una validación cruzada de 20 particiones con una proporción del 60 %-20 %-20 % para el conjunto de entrenamiento, validación y prueba, respectivamente, este último conjunto siempre balanceado entre hombres y mujeres. Las pruebas realizadas reflejan el mismo comportamiento observado en los casos sintéticos (comparar Fig. 1 b y Fig. 2 a), donde el mejor

caso posible para la población completa se da cuando la misma se encuentra balanceada. Los resultados de la validación cruzada demuestran que una desviación sobre la distribución 50-50 se traduce en un aumento del error (Fig. 2 b).

3. Discusión y trabajo futuro

Se estudió primero el efecto del desbalance de un atributo protegido presente en una base de datos sintética sobre el desempeño de un modelo en distintas subpoblaciones, observando que tanto la población menos representada como la población completa se ven afectadas por dicho grado de desbalance. Este ejemplo de juguete permitió generar intuiciones que fueron luego confirmadas en la base de datos real de IMDb, la cual evidenció la misma dinámica. Esos resultados son el primer paso de un trabajo que busca comprender si este comportamiento se manifiesta en otras bases de datos reales, como UTK [9] donde, además del género, se considerará el atributo protegido del color de piel de cada individuo, para luego diseñar algoritmos que puedan atenuar este problema, agregando a la función de costo penalizadores que des-incentiven sesgos. Se evaluará en este sentido el uso de estrategias de optimización basadas en relajación Lagrangeana[8].

Referencias

1. Escalas, E., Echeveste, R., Peterson, V., Ferrante, E.: Desbalance de datos en términos de atributos protegidos: análisis de su impacto en un clasificador lineal (2020)
2. Hutson, M., et al.: Even artificial intelligence can acquire biases against race and gender. Science Magazine 10 (2017)
3. 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. Proceedings of the National Academy of Sciences (2020)
4. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) 54(6), 1–35 (2021)
5. Prates, M.O., Avelar, P.H., Lamb, L.C.: Assessing gender bias in machine translation: a case study with google translate. Neural Comp and App (2019)
6. Rothe, R., Timofte, R., Gool, L.V.: Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision 126(2-4), 144–157 (2018)
7. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision 115(3), 211–252 (2015)
8. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv preprint arXiv:1707.09457 (2017)
9. Zhifei, Z., Yang, S., Hairong, Q.: Age progression/regression by conditional adversarial autoencoder. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2017)
10. Zou, J., Schiebinger, L.: AI can be sexist and racist—it's time to make it fair. Nature 559, 324–326 (2018)