Desbalance de datos en términos de atributos protegidos: análisis de su impacto en un clasificador lineal

Eugenia Escalas[1], Rodrigo Echeveste[1]*, Victoria Peterson[2]*, Enzo Ferrante[1]*

[1] Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), FICH-UNL/CONICET, Argentina
[2] Instituto de Matemática Aplicada del Litoral, IMAL, UNL/CONICET, Argentina

Abstract. En este trabajo se busca estudiar el impacto del desbalance en los datos utilizados para entrenar un clasificador lineal, centrando el análisis en atributos protegidos. Dichos atributos, tales como género, grupo étnico o edad, no constituyen la clase objetivo del clasificador, sino que corresponden a características demográficas que pueden ser o no parte del problema a resolver. Los resultados obtenidos mediante experimentos sintéticos simples muestran que la exactitud sobre una población dada se deteriora cuando se encuentra subrepresentada en el conjunto de datos de entrenamiento. En todos los casos, el rendimiento del clasificador sobre la población completa es máximo cuando este conjunto de datos se encuentra balanceado en lo que respecta a atributos protegidos. Estas conclusiones son el primer paso de un trabajo que busca mostrar cómo puede atenuarse este inconveniente incorporando penalizantes que desincentiven un aumento de la exactitud sobre un subconjunto de la población en desmedro de otra.

1 Introducción

La rápida evolución de los algoritmos de inteligencia artificial (IA) en diferentes áreas y aplicaciones ha llevado a que la opinión y el comportamiento de la sociedad estén cada vez más influenciados por estas tecnologías. Las noticias que leemos, el camino que tomamos para volver a casa e incluso la forma en que diagnosticamos enfermedades, son sólo algunos ejemplos del uso de IA en nuestra vida cotidiana. En este contexto cobra especial relevancia el estudio de los posibles riegos y fallas asociados a estos modelos. En efecto, existe evidencia de que ciertos sesgos sociales, han sido no sólo heredados por los sistemas de IA, sino también amplificados en múltiples contextos [10,4,1]. Tal es el caso, por citar un ejemplo, de un sistema de reconocimiento automático de imágenes, que al analizar la fotografía con un hombre en una cocina, automáticamente categoriza a ese individuo como *mujer* [9].

Si bien el concepto de "justicia" (o *fairness* en inglés) en IA aún no tiene una definición unívoca [8], cada vez son más los grupos de investigación que invierten esfuerzos en evitar los potenciales daños que la falta de justicia y equidad

* Coautoría en partes iguales

en los algoritmos de IA podría causar. En particular, la comunidad científica de "innovaciones en género" (*gendered innovations*) aboga, desde hace varios años, por la incorporación de la dimensión sexo-género en el diseño experimental y el análisis de los desarrollos científico-tecnológicos [7], entre los cuales se incluyen los modelos de IA. Si bien tanto el modelo de predicción como los datos pueden ser las fuentes del sesgo en IA, diversos trabajos publicados recientemente [6,2,3,5] proveen evidencia de que el desbalance de género en las bases de datos impacta directamente en el rendimiento de los clasificadores, presentando menor rendimiento en los grupos subrepresentados.

En este trabajo buscamos caracterizar el impacto que conlleva el desbalance en los datos de entrenamiento mediante ejemplos sintéticos simplificados. Si bien la mayor parte de la literatura sobre aprendizaje automático con datos desbalanceados centra su atención en el desbalance dado por las clases de interés (*target*), en este trabajo nos centraremos en el desbalance sobre atributos protegidos como el género, grupo étnico o edad, que no constituyen la clase *target* del problema, sino una característica más de los datos.

2 Formulación del Problema

En esta primera etapa estudiamos un proceso de clasificación binaria (ejemplificado con pacientes sanos o enfermos) considerando una base de datos sintética muy simplificada (un problema de juguete). El objetivo es estudiar el impacto que genera un desbalance en la representatividad de distintos subgrupos de la población. En particular, nos centraremos en estudiar cómo afecta el desbalance en los datos de entrenamiento a la exactitud de un clasificador, tanto de los grupos menos representados, como de la población completa.

Denotemos por X el espacio de entrada, donde cada $x \in \mathbb{R}^2$ representa las características de cada individuo a clasificar, por Y el conjunto de las etiquetas (en $\{0, 1\}$, según el paciente esté sano o enfermo), y Z una característica protegida que indica a qué subgrupo de la población pertence cada individuo, que indicamos aquí como H (hombre) o M (mujer). Este atributo protegido no es otorgado en forma explícita al clasificador pero sí determina la distribución de los datos de entrada para ese subgrupo.

Asumimos que la distribución real de los datos en el problema está descripta por cuatro distribuciones normales multivariadas:

$$\mathcal{P}\left(x \,|\, y = 0/1, z = H/M\right) = \mathcal{N}(\boldsymbol{\mu}^{0/1}_{H/M}, \boldsymbol{\Sigma}). \tag{1}$$

Donde cada x del espacio de entrada proviene de una de las cuatro distribuciones normales, según el individuo sea mujer u hombre, y según sea un sujeto sano o enfermo. Para simplificar el problema aún más, hemos supuesto que las cuatro distribuciones sólo difieren en la media ($\boldsymbol{\mu}$), pero no en la matriz de covarianza ($\boldsymbol{\Sigma}$). Asumimos además que en la población real, cada uno de estos cuatro subgrupos se encuentran igualmente representados, es decir que la fracción de mujeres puede describirse como $\mathcal{P}(z = M) = f_M = 0.5$ y la fracción de pacientes

sanos como $\mathcal{P}(y=0) = f_S = 0.5$. Consideramos, sin embargo, que en el conjunto de entrenamiento contamos con un desbalance de género, dado por $\tilde{f}_M \in [0,1]$, cuyo valor variaremos sistemáticamente.

3 Resultados

Entrenamos un clasificador lineal para distintas configuraciones de las medias $\boldsymbol{\mu}_{H/M}^{0/1}$ (ver filas de la Fig. 1) y distinto grado de desbalance de género en el conjunto de entrenamiento, dado por $\tilde{f}_M$ (columnas de la Fig. 1 a, y abscisas en la Fig. 1 b). El clasificador utilizado fue un *Linear SVM*, empleando la función *SGDClassifer* de la librería *sklearn*, con función de perdida *hinge*. La tolerancia establecida fue de 1×10^{-4}. Los experimentos numéricos se repitieron $N_{iter} = 100$ veces, para cada valor de $\tilde{f}_M$, generando en cada caso una muestra aleatoria de $n = 250$ puntos de cada distribución, para formar el conjunto de entrenamiento. Luego procedimos a calcular la exactitud sobre un conjunto de test balanceado ($f_M = 0.5$, Fig. 1 b). Se presenta la exactitud promedio ($\pm$ una desviación estándar) calculada sobre las múltiples repeticiones, para cada caso, y en función del grado de desbalance de género en el conjunto de entrenamiento.

Observamos primeramente que la exactitud sobre una subpoblación dada se deteriora cuando la misma se encuentra poco representada en el conjunto de entrenamiento. Como es de esperar, este deterioro es más marcado cuánto mayor sea la "tensión" entre los dos grupos en términos de sus respectivas fronteras de decisión óptimas (comparar casos 1 a 3). Lo interesante es que no sólo la exactitud sobre la población menos representada es la que sufre. En las curvas de exactitud sobre la población total (líneas verdes en la Fig. 1 b), vemos que en todos los casos el rendimiento del clasificador es máximo cuando el conjunto de entrenamiento está balanceado.

4 Discusión y trabajo futuro

En el presente trabajo presentamos un problema de juguete que nos permite ilustrar cómo el desbalance en términos de un atributo protegido (por ejemplo, el género) presente en una base de datos usada para entrenar un clasificador puede producir una pérdida en el rendimiento de dicho modelo a la hora de emplearlo sobre la población real (si la misma está balanceada). En este sentido, es importante destacar que según este modelo simple, no es solamente la población menos representada la que se ve perjudicada (un resultado lamentable, pero esperable), sino que la exactitud del clasificador sobre el conjunto total de la población también disminuye cuando el conjunto de entrenamiento está desbalanceado.

Si bien los resultados aquí expuestos corresponden a una distribución particular de los datos, y a un clasificador lineal, es de esperar que la tendencia de encontrar un máximo en la exactitud del clasificador sobre la población total cuando los subgrupos están balanceados tenga un carácter más general. Esto se debe a la llamada "ley de los rendimientos decrecientes", que indica que para

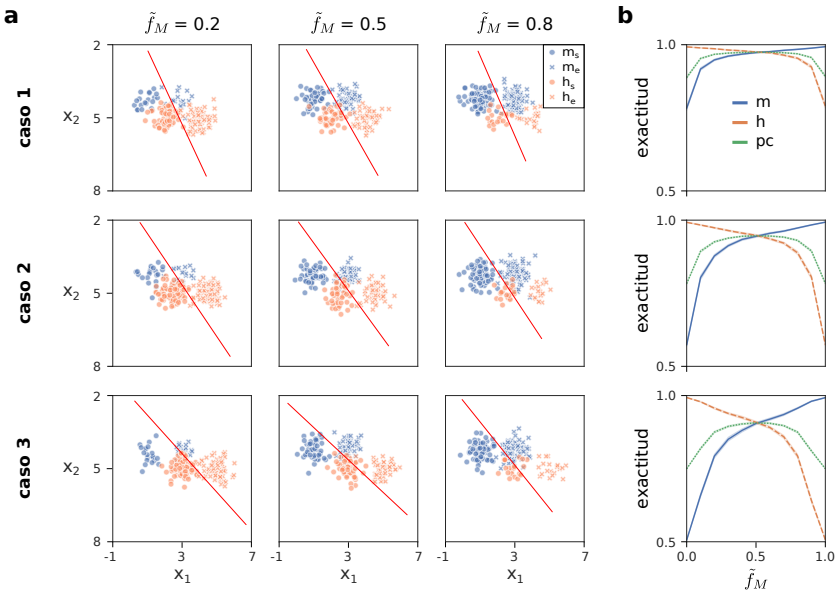

Fig. 1: **Exactitud de un clasificador lineal según el grado de desbalance de género en el conjunto de entrenamiento**
a, Datos de entrenamiento (puntos) y frontera de decisión del clasificador lineal entrenado (línea roja), en tres casos distintos (filas). Los puntos azules corresponden a las mujeres, y los naranjas a los hombres. Los pacientes sanos se indican con un círculo, y los enfermos con una cruz. **b**, Exactitud del clasificador entrenado, evaluada sobre el conjunto de test, en función del desbalance de género del conjunto de entrenamiento ($\tilde{f}_M$). Nota: el conjunto de test está balanceado ($f_M = 0.5$). Se muestran los resultados sobre cada población individual (mujeres en azul y hombres en naranja), así como en la población completa (en verde). Se presenta la curva promedio sobre las distintas iteraciones ($\pm$ una desviación estándar).

una población individual es de esperar que la exactitud de un clasificador se incremente al incorporar más puntos, pero con una pendiente progresivamente menor, siempre y cuando los datos se incorporen de forma aleatoria y provengan de la misma distribución (ver líneas azules en Fig. 1 b). Notando que la curva para la otra población (en naranja) es una versión espejada de la primera, y que la exactitud sobre la población total (en verde) es el promedio sobre las exactitudes individuales, resulta claro que este fenómeno será tan general como lo sea la ley de los rendimientos decrecientes.

Estos resultados son el primer paso de un trabajo que busca mostrar cómo puede atenuarse este problema agregando a la función de costo un penalizador que des-incentive un aumento de la exactitud sobre un subconjunto de la población en desmedro de la otra. Considerando que pueden existir más de dos sub-grupos o clusters en la población, y que la información sobre el sub-grupo de pertenencia de un dado sujeto puede no estar disponible, se evalúa utilizar criterios tanto supervisados como no supervisados para identificar sub-poblaciones (o minorías) en los datos como un paso previo al entrenamiento de los modelos. Dichos experimentos numéricos se encuentran en este momento todavía en progreso.

Finalmente se buscará extender estos resultados utilizando bases de datos reales, para confirmar o descartar las intuiciones generadas a partir de bases de datos sintéticas.

References

1. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Advances in neural information processing systems. pp. 4349–4357 (2016)
2. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. pp. 77–91 (2018)
3. Chapman, K.R., Tashkin, D.P., Pye, D.J.: Gender bias in the diagnosis of copd. Chest 119(6), 1691–1695 (2001)
4. Hutson, M., et al.: Even artificial intelligence can acquire biases against race and gender. Science Magazine 10 (2017)
5. 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)
6. 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)
7. Schiebinger, L., Schraudner, M.: Interdisciplinary approaches to achieving gendered innovations in science, medicine, and engineering1. Interdisciplinary Science Reviews 36(2), 154–167 (2011)
8. Verma, S., Rubin, J.: Fairness definitions explained. In: 2018 IEEE/ACM International Workshop on Software Fairness (FairWare). pp. 1–7. IEEE (2018)
9. 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)
10. Zou, J., Schiebinger, L.: AI can be sexist and racist—it's time to make it fair. Nature 559, 324–326 (2018)