IDEAS EN CIENCIAS DE LA INGENIERÍA
Facultad de Ingeniería, Universidad Autónoma del Estado de México

Comportamiento de algoritmos de sobre-muestreo en Big Data

Behavior of over-sampling algorithms in Big Data

A. Guzmán-Ponce[1], César Ferri[2], J. S. Sánchez-Garreta[1], J. R. Marcial-Romero[3]

Recibido: 17 de noviembre de 2021 – Aceptado: 4 de abril de 2022

RESUMEN

El desbalance de clases es una de las complejidades de los datos ampliamente estudiada en el campo de la ciencia de datos. A menudo dificulta el proceso de extracción de conocimiento, sesgando el aprendizaje hacia instancias de clase mayoritaria. La creciente generación de datos que estamos viviendo agrava el escenario anterior. Los desafíos en *Big Data* implica la necesidad de adaptar o crear nuevas técnicas para las restricciones de escalabilidad, dando lugar al desarrollo de técnicas que solventen el desbalance de clases en grandes volúmenes de datos, siendo la mayoría de estas basadas en el algoritmo SMOTE, en razón de tener un mejor desempeño en conjuntos "pequeños". En este trabajo realizamos un análisis del comportamiento de los métodos de sobre-muestreo en *Big Data*, a través de medidas de complejidad que permiten conocer las características de los conjuntos de datos procesados. Los resultados obtenidos corroboran que el problema de desbalance de clases en *Big Data* no es el único problema que debe abordarse; por otro lado, el comportamiento de SMOTE en *Big Data* no

[1] Department of Computer Languages and Systems, Institute of New Imaging Technologies, Universitat Jaume I, 12071 Castelló de la Plana, Spain
[2] Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Spain
[3] Facultad de Ingeniería, Universidad Autónoma del Estado de México, Toluca, México.
Autor de correspondencia: aguzman@uji.es

es comparable al logrado en conjuntos de datos pequeños, debido a la presencia de redundancia por parte del proceso de interpolación.

PALABRAS CLAVE: sobre-muestreo; SMOTE; Big Data

ABSTRACT

Class imbalance is one of the data complexities widely studied in the field of classification. It often hinders the modelling process, biasing the learning towards majority-class examples. The fast-growing generation of data that we are experiencing, aggravates the former scenario. In Big Data imbalanced classification, the challenges imply both addressing the skewed class distribution, and 4the need to adapt or create new techniques for scalability constraints. Thus, becoming the development of techniques that handled the class imbalance on a high amount of data, as most of these are based on the original SMOTE algorithm, by virtue of their better performance on ``small" data sets. In this paper, we performed an analysis of the behaviour of over-sampling methods on Big Data, through complexity measures which allow knowing the characteristics of the processed data sets. The obtained results corroborated that the class imbalance problem on Big Data sets is not a unique problem that must be addressed; on the other hand, the behaviour for SMOTE on Big Data is not comparable to that achieved in “small data" problems, due to the presence of redundancy by the interpolation process

KEYWORDS: over-sampling; SMOTE; Big Data.

INTRODUCCIÓN

En la actualidad la recolección y el análisis de datos es de suma importancia para empresas y organizaciones, convirtiendo a los datos en un activo fundamental para la toma de decisiones. Para lograr este objetivo, los datos son usados para obtener patrones de los datos y generar un modelo que contiene el conocimiento extraído (García et al., 2020).

Con el incremento de datos provenientes de diversas fuentes con diferentes tipos de datos, a la capacidad de recopilarlos y procesarlos de manera adecuada se le conoce como *Big Data* (Luengo et al., 2020). Este concepto requiere de atención en el análisis de datos, dado que por su naturaleza los datos se encuentran bajo peculiaridades que decrementan el rendimiento de un clasificador, es decir, este tipo de algoritmos se ven afectados por la calidad de los datos. En *Big Data*, el término *Smart Data* hace referencia a datos de calidad suficiente para lograr modelos de alto rendimiento.

En una gran variedad de problemas del mundo real como nuevos descubrimientos físicos, diagnóstico de enfermedades raras, detección de fraude electrónico, entre otros, comparten la particularidad de no ser iguales en términos de tamaño de datos por clase, lo que se conduce a un escenario de desbalance (García et al., 2018). Esta situación sucede cuando una de las clases tiene un número de instancias significativamente mayor en comparación con el resto de las clases. En general, en un conjunto de dos clases, a la clase menos representada se conoce como positiva, denotada por C^+, mientras que la clase más representada se conoce como negativa denotada por C^- (Basgall et al., 2019) (Figura 1).

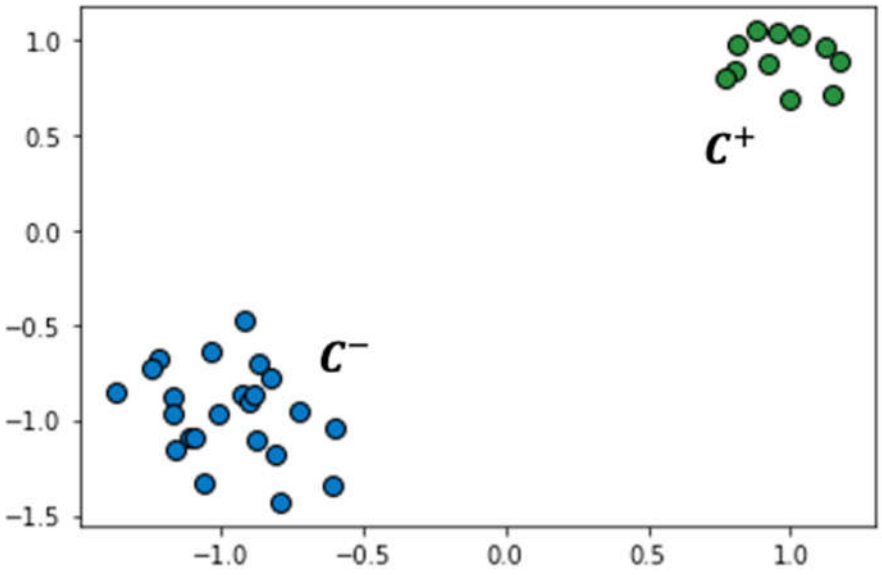

FIGURA 1.

CONJUNTO DE DATOS DESBALANCEADO

Una manera de abordar el problema de desbalance de clases es equilibrar la distribución de clases volviendo a muestrear el espacio de datos (S. García et al., 2016, Rendón et al., 2020), estas estrategias de re-muestreo son independientes del clasificador y sólo trabajan sobre el conjunto de datos, suelen dividirse en dos técnicas: sobre-muestreo y bajo-muestreo. La primera crea instancias de la clase positiva, mientras que la segunda técnica elimina instancias de la clase negativa.

El uso de métodos de sobre-muestreo se debe principalmente a disminuir la pérdida de datos en clase positiva que pueden ser útiles para generar conjuntos de datos de alta calidad, esto sucede comúnmente en escenarios con conjuntos de datos pequeños (*Small Data*) (V. García et al., 2020). Un conjunto de datos pequeños se caracteriza por su volumen generalmente limitado, siendo menor a Gigas de información, sin recopilación continua y con una variedad limitada de tipos de datos (Kitchin & Lauriault, 2015).

Sin embargo, en términos de grandes volúmenes de datos surge la cuestión ¿El sobre-muestreo mejora la calidad de los datos en escenarios de *Big Data*? Para responder esta cuestión, es necesario conocer las propiedades de los conjuntos de datos previas al tratamiento del desbalance de clases y posterior análisis una vez aplicada las técnicas de sobre-muestreo.

ESTRATEGIAS DE SOBRE-MUESTREO

A lo largo del tiempo se han propuesto un gran número de estrategias para hacer frente al desbalance de clases en escenarios *Small Data*. Sin embargo, para *Big Data* algunas investigaciones se han llevado a cabo mediante el escalado de métodos de sobre-muestreo bien conocidos (del Río et al., 2014).

El método clásico de sobre-muestreo es aleatorio (ROS) (Batista et al., 2004) el cual replica instancias tomadas aleatoriamente de la clase positiva hasta que el tamaño de la clase positiva es igual al de la clase negativa. Además de ROS, SMOTE (Gutiérrez et al., 2017) ha sido ampliamente usado, ya que funciona creando instancias a través de la interpolación de cada instancia positiva con sus vecinos.

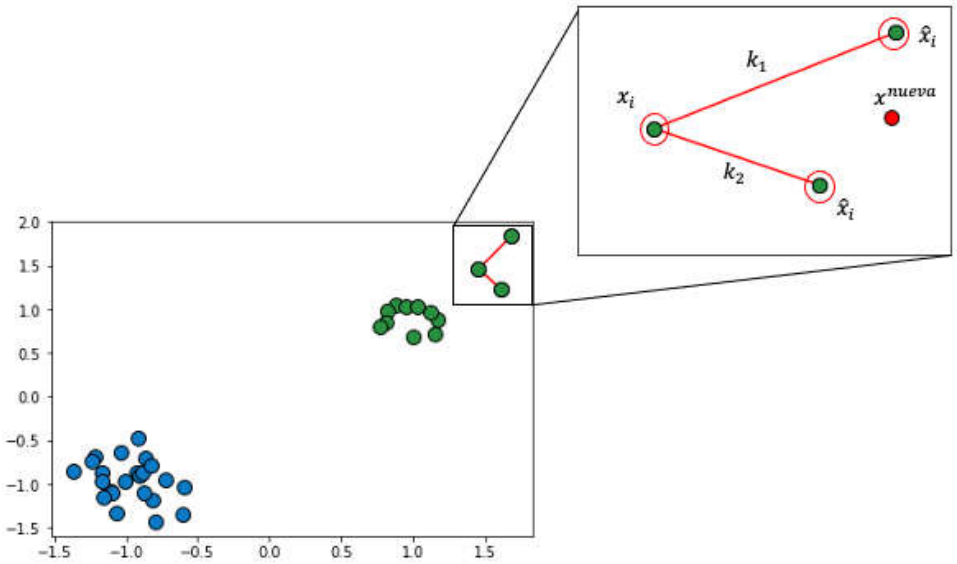

FIGURA 2.
REPRESENTACIÓN DE SMOTE

El algoritmo de SMOTE toma instancias de clase positiva y a través de sus vecinos más próximos genera nuevas instancias por interpolación. Del ejemplo de la Figura 2, la instancia x^{nueva}, se genera de combinar las características de la instancia x_i con características de sus vecinos.

En torno al método SMOTE también hay algunas propuestas que versan en el procedimiento para abordar el conjuto de datos, es decir, de manera local o global. SMOTE-MR (Basgall et al., 2018), es una versión local de SMOTE, donde para calcular los vecinos k de una instancia de la clase positiva, SMOTE-MR utiliza la misma partición a la que pertenece una instancia. Por el contrario, SMOTE-BD (Basgall et al., 2019) se desarrolla como una técnica global, donde el cálculo de k vecinos más cercanos (kNN) se basó en el método KNN-IS (Maillo et al., 2017), este algoritmo a través de múltiples reductores determina cuáles son los k vecinos más cercanos finales de cada partición.

MÉTRICAS DE CALIDAD DE LOS DATOS

Además del problema de desbalance de clases, existen otros factores que inciden en el rendimiento del clasificador, denominadas complejidades de los datos. El traslape de clases, implica regiones ambiguas del espacio de características donde la probabilidad de las clases es similar. Otro factor negativo es el ruido, lo que implica instancias mal etiquetas, que pueden resultar en decisiones incorrectas por parte del clasificador. Con los grandes cúmulos de datos, el número de características excede significativamente el número de instancias del conjunto, lo que provoca el problema conocido como alta dimensionalidad. La Figura 3 representa estos problemas en un conjunto de dos clases.

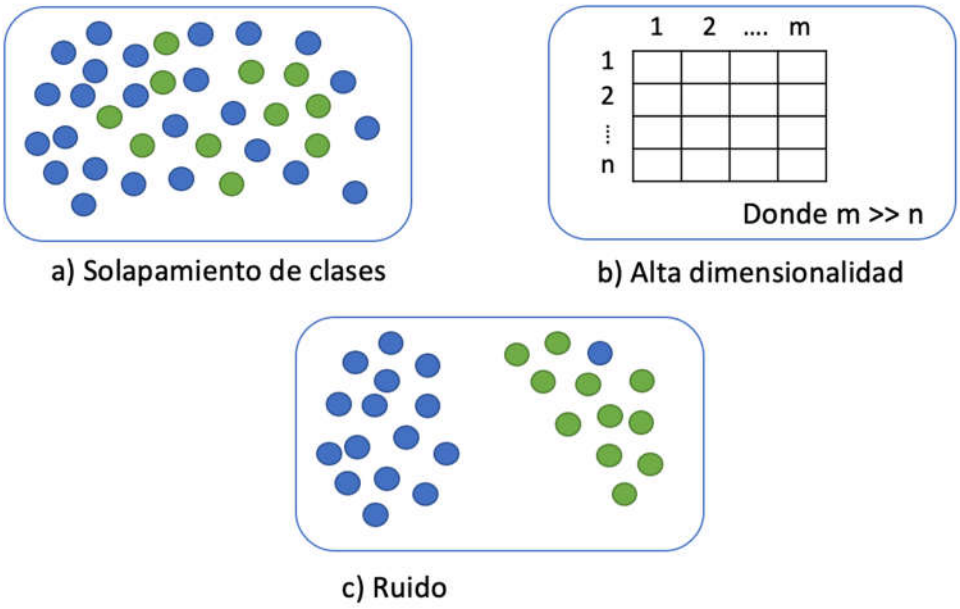

FIGURA 3.

COMPLEJIDADES DE LOS DATOS

Comúnmente el comportamiento que un conjunto de datos tiene en un clasificador es medido por el rendimiento del propio algoritmo. Sin embargo, existen algunas métricas que permiten determinar la calidad del conjunto de datos independientemente del clasificador.

Considere que un conjunto de datos CD esta conformado por n instancias, y cada instancia está compuesta por (x, y) atributos o características descritas como un arreglo $x = [x_1, \quad , x_m]$ de m características. La variable y está compuesta por n_c clases. Para determinar la presencia de traslape de clases la métrica comúnmente usada es el radio discriminante de Fisher ($F1$), (Maillo et al., 2020) escala esta métrica en escenarios de *Big Data*, la cual se define por la ecuación 1:

$$F1 = máx_{i=1}^{m} \frac{\sum_{j=1}^{n_c} n_{cj}\left(\mu_{cj}^{fi} \quad \mu^{fi}\right)^2}{\sum_{j=i}^{n_c}\sum_{l=1}^{n_{cj}}\left(x_{li}^{j} \quad \mu_{cj}^{fi}\right)^2} \tag{1}$$

Donde n_{cj} es el número de instancias de la clase j, μ_{cj}^{fi} es el promedio de la i-ésima característica de las instancias de la clase j, μ^{fi} es el promedio de la i-ésima característica de todas las instancias y x_{li}^{j} es el valor especifico de la i-ésima característica para una instancia particular x.

Por otro lado, para tener una descripción general de cómo funcionan las características en conjunto, se usa la Eficiencia de funciones colectivas (F4). En general, cuenta el número de instancias afectadas considerando todas las características en un proceso iterativo. Esta definida por la ecuación 2:

$$F4 = \frac{n \quad n_o\left(f_{máx}(T_l)\right)}{n} \tag{2}$$

El proceso iterativo de $f_{\text{máx}}(T_l)$ se define por la ecuación 3:

$$f_{máx}(T_i) = \left\{f_j \middle| \, máx_{j=1}^{m} \quad n_0\left(f_j\right)\right\} \tag{3}$$

METODOLOGÍA

En esta sección, se describe el estudio realizado para la comparación experimental del desempeño de los métodos para el manejo del desbalance de clases en *Big Data*. En general, la Figura 4 describe la Metodología seguida.

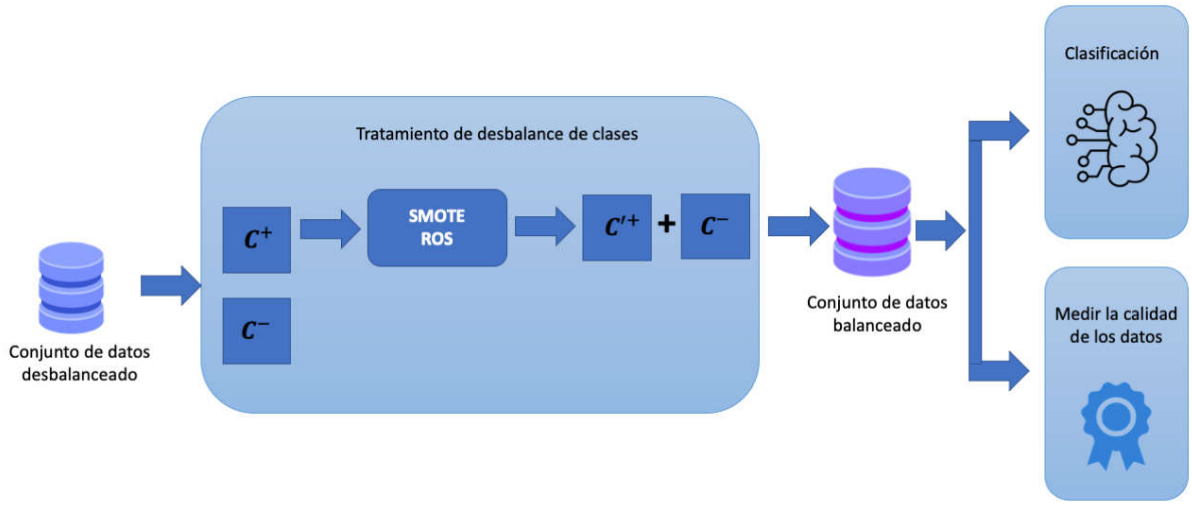

FIGURA 4.

METODOLOGÍA

En general, para grandes conjuntos de dos clases desbalanceados, se aplican las técnicas de sobre-muestreo. Una vez tratado el desbalance se clases, se evalúa el rendimiento del clasificador, además de evaluar la calidad de los conjuntos de datos generados.

CONJUNTO DE DATOS

Basamos la experimentación en 6 conjuntos de *Big Data* y 6 conjuntos de *Small Data*, todos ellos de dos clases con desbalance, tomados del repositorio UCI *Machine Learning* (Lichman, s. f.). Sus detalles se describen en la Tabla 1. Los conjuntos fueron seleccionados en función a su grado de desbalance (IR), es decir, la relación entre el número de instancias de la clase negativa y el número de instancias de la clase positiva. El grado de desbalance de estos conjuntos varía entre 2.57 y 40.

TABLA 1.
CONJUNTOS DE DATOS DESBALANCEADOS USADOS PARA BIG DATA Y SMALL DATA

	Conjunto	Total	Número de Atributos	Grado de desbalance (IR)
1	MiniBooNE	29178-74870	49	2.57
2	Susy	542434-2169739	18	4.00
3	poker	39062-410960	10	10.52
4	HEPMASS	262435-4200169	28	16.00
5	HIGGS	291417- 4663335	28	16.00
6	Covtype	16256 - 448421	54	27.58
7	vowel0	90 - 898	13	9.98
8	glass2	17 - 197	9	11.59
9	cleveland	13 - 160	13	12.31
10	glass4	13 - 201	9	15.47
11	yeast5	44 - 1440	8	32.73
12	abalone	58 - 2280	8	39.31

Se usa la validación cruzada en 5 particiones, donde el 80% de las instancias se dedicaron a entrenamiento y el 20% restante a pruebas. La Tabla 2 muestra el promedio de ejecutar los algoritmos.

TRATAMIENTO DE DESBALANCE

Con el fin de comparar el comportamiento de rendimiento de los métodos de sobre-muestreo, el estudio experimental se aplicarán los métodos de balanceo descritos en la Sección de Estrategias de sobre-muestreo, estos métodos son ROS, SMOTE-MR (*S-MR*), SMOTE-BD (*S-DB*). Sin embargo, la mayoría necesita algunos parámetros libres, por lo que las especificaciones técnicas del método de sobre-muestreo son: el porcentaje promedio de sobre-muestreo fue de *100*, usamos particiones de *128*, los k vecinos más cercanos para SMOTE fueron de *5*, usando la distancia euclidiana.

MÉTRICAS DE EVALUACIÓN

Con el fin de describir el rendimiento del clasificador, se ejecutó el árbol de decisión de la *API MLlib Spark*. Para un problema de dos clases, la matriz de confusión es usada para obtener las métricas de calidad. Así, las métricas que miden el rendimiento están dadas por la cantidad de instancias de cada clase que se clasificaron correctamente, nombradas como verdadero-positivo (TP) y verdadero-negativo (TN), mientras que la cantidad de casos positivos y negativos las instancias negativas mal clasificadas se denominan falso-positivo (FP) *y falso-negativo* (FN), respectivamente. Para determinar el rendimiento del algoritmo de aprendizaje en conjuntos de datos desbalanceados, comúnmente se emplea la media geométrica (Ecuación 4)

$$G_{mean} = \sqrt{\left(TP/(TP+FN)\right) \quad \left(TN\ /\ (TN+FP)\right)} \tag{4}$$

Se evalúa la calidad de los datos mediante la determinación de redundancia (métrica F1) y la eficiencia de características (métrica F4).

INFRAESTRUCTURA

Manejar enormes cantidades de datos requiere procesamiento de alto rendimiento y potencia de almacenamiento. Todos los experimentos se han llevado a cabo en el servidor *falco* del grupo de trabajo VRAIN, con un microprocesador Intel(R) Core(TM) i9-10920X CPU @ 3.50GHz y 132 GB de RAM. En cuanto al software se utilizó Apache Spark 2.2.0.

RESULTADOS

El objetivo de las técnicas de sobre-muestreo es disminuir el sesgo de aprendizaje, mediante la creación de nuevas instancias. La Tabla 2 reporta la media geométrica promedio del esquema de validación cruzada de cinco veces cada conjunto de datos, por método de sobre-muestreo. Adicionalmente, se proporciona el rango promedio de *Friedman* para pequeños y

grandes conjuntos. Como base de comparación, se incluyen los resultados de los conjuntos de datos sin tratar el desbalance.

TABLA 2.
RESULTADOS OBTENIDOS CON LOS ALGORITMOS DE SOBRE-MUESTREO

Conjunto	**Referencia**	**ROS**	**S-MR**	**S-DB**
MiniBooNE	85.2	87.9	87.9	**88.1**
Susy	68.8	**76.8**	76.3	76.3
Poker	17.0	58.1	**59.7**	53.9
HEPMASS	71.9	**83.3**	66.8	65.5
HIGGS	11.6	**66.0**	64.4	64.9
Covtype	72.8	**93.2**	92.6	93.0
Promedio de rank	3.7	**1.4**	2.5	2.4
Vowel0	92.3	95.3	96.2	**97**
Glass2	28.9	51.5	**77.1**	70.6
Cleveland	41.1	51.4	**75.6**	75
Glass4	83.1	85.4	**88.7**	88.7
Yeast5	82.1	**92.1**	89.3	91.8
Abalone	22	80.3	79.4	**80.4**
Promedio de rank	4	2.5	1.9	**1.6**

Los resultados de la Tabla 2 señalan que para conjuntos de *Big Data*, ROS tiene un mejor rendimiento al tener el mejor rango promedio de Friedman en comparación con las propuestas basadas en SMOTE. Este comportamiento es diferente en conjuntos de datos pequeños, como se observa, el rango promedio obtenido por SMOTE-DB es el mejor.

Centrándose en el conjunto poker, el bajo rendimiento en SMOTE sugiere que es causado por el procedimiento de interpolación, dado que SMOTE crea instancias sintéticas enfocadas en el espacio de características, y este conjunto de datos tienen características

discretas, las nuevas instancias no brindan suficiente conocimiento sobre los grandes conjuntos de datos.

La medida F1 determinar si las clases están superpuestas, en la Figura 5 se observa, que los valores más bajos se presentan en los conjuntos de referencia, lo que incurre en traslape de clases. De manera especifica, para conjuntos de datos pequeños con técnicas de SMOTE, se observa que los valores obtenidos son mayores. En contraste con los conjuntos de *Big Data*, en todos los métodos de sobre-muestreo la medida F1 se mantiene bien ajustada y el valor resultante es bajo, lo que sugiere que la complejidad incurre en la existencia de redundancia, aunque se generen nuevas instancias por interpolación.

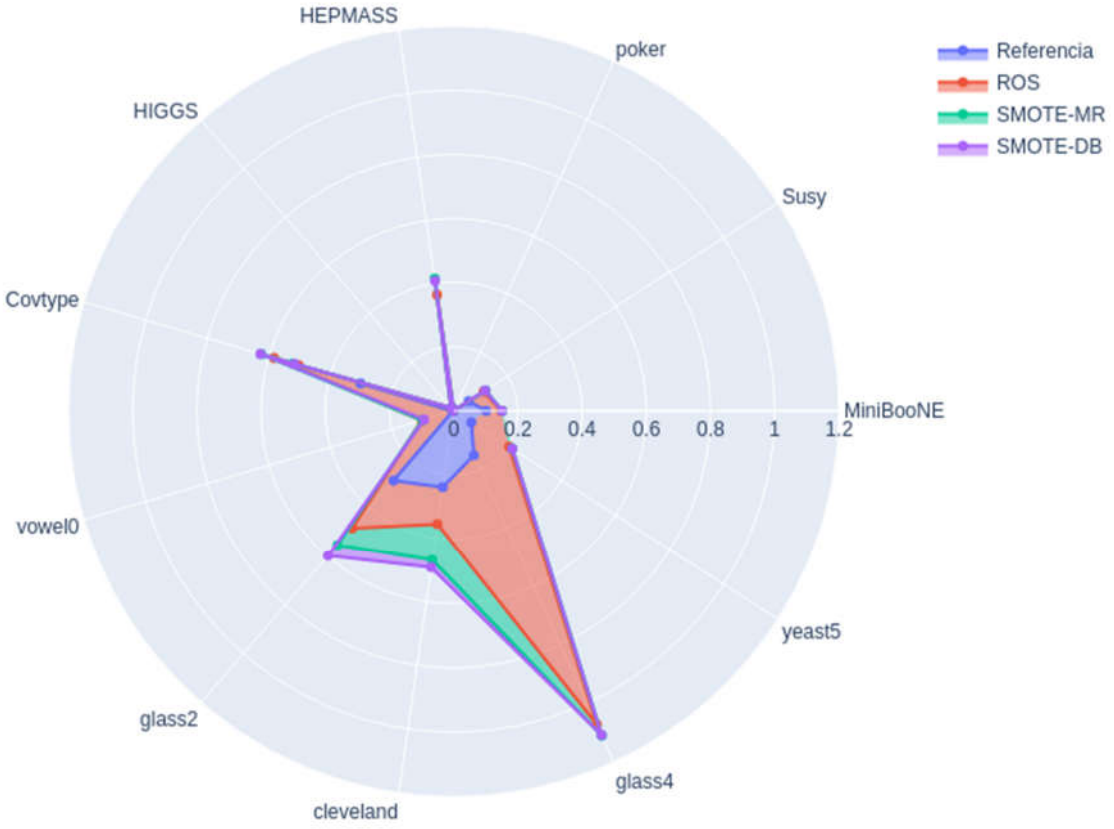

FIGURA 5.

RELACIÓN MÁXIMA DE DISCRIMINACIÓN (MÉTRICA F1)

De hecho, estos resultados evidencian que el uso de la distancia euclidiana puede no ser una norma adecuada para los conjuntos de datos utilizados, debido a la interpolación. El término de proximidad se define como todas las instancias son aproximadamente equidistantes entre sí, y la distancia euclidiana asume que todos los atributos son igualmente importantes (Maldonado et al., 2019), en este sentido, la similitud del coseno puede conducir a un mejor resultado debido a algunas propiedades de las instancias que hacen que los pesos sean mayores sin diferencia.

Podemos destacar que para conjuntos de *Big Data*, un método de sobre-muestreo basado en el espacio de características no conduce a mejores resultados, de hecho, es necesario considerar el espacio de datos en general. Dado que manejamos una gran cantidad de datos, seguimos enfrentando problemas como el traslape de clases, donde es evidente que la frontera de solución no está clara dada la redundancia de instancias que puede tener *Big Data.*

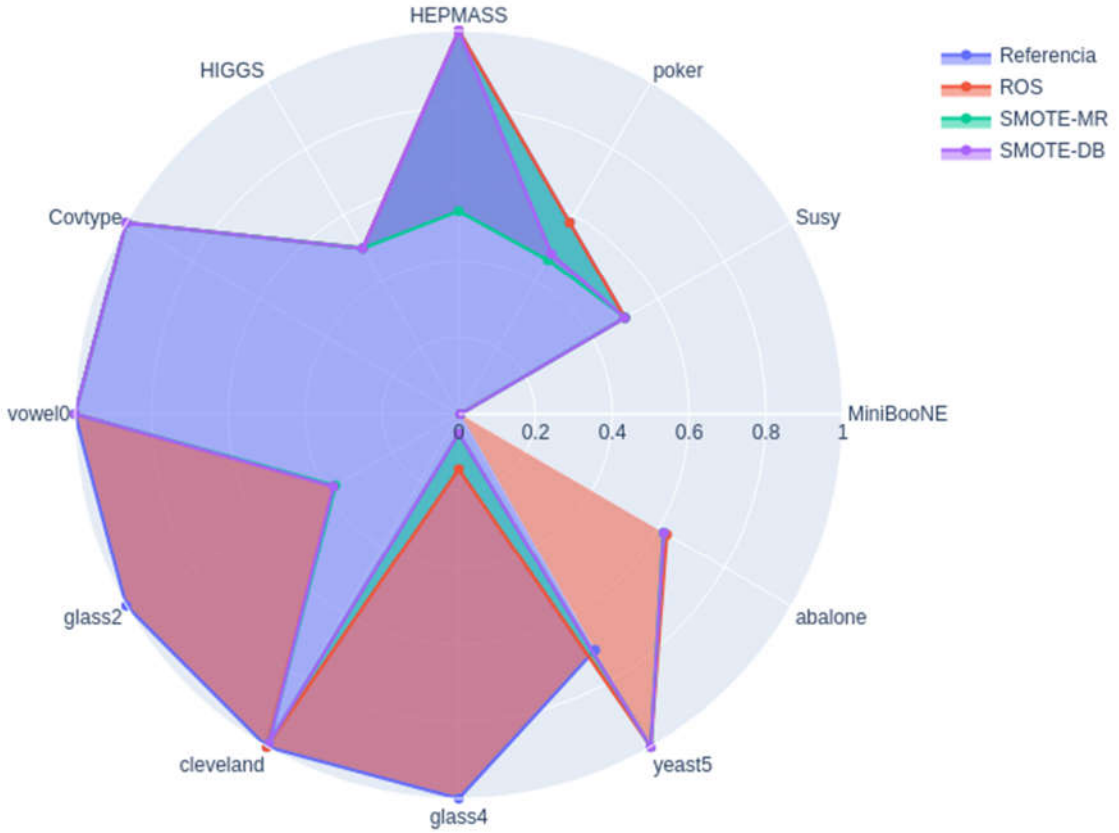

FIGURA 6.

EFICIENCIA DE CARACTERÍSTICA COLECTIVA (MÉTRICA F4)

La medida F4 obtiene una descripción general de cómo funcionan juntas las características. De este modo, los valores más bajos de F4 indican que es posible discriminar más instancias y, por lo tanto, que el problema es más simple. De la Figura 6, los valores obtenidos por el conjunto base, podemos deducir que los conjuntos de datos en ambos escenarios (*Big Data* y *Small Data*) no es posible discriminar instancias, debido a que todas ellas obtuvieron los valores más altos en la medida F4. Para *glass4* las técnicas de SMOTE reducen la complejidad, este comportamiento sugiere que la generación de instancias sintéticas mantiene características discriminatorias de las originales, no así para el método aleatorio.

Los resultados obtenidos en ambos escenarios sugieren que, en contexto de *Big Data*, las características intrínsecas de los datos se conservan. En la mayoría de los conjuntos de datos aplicando SMOTE sufren de traslape de clases, debido a la distribución de las instancias sintéticas. De hecho, esto expondría que el espacio de muestra en los conjuntos de *Big Data* es aparentemente pequeño en comparación con la cantidad de instancias en cada conjunto de datos.

CONCLUSIONES

El desbalance de clases sigue siendo uno de los problemas más comunes y relevantes, no sólo en conjuntos de datos "pequeños" sino también en conjuntos de *Big Data*. Cabe señalar que en este último caso se agrava especialmente ya que se trata de una gran cantidad de datos generado con frecuencia. Realizamos una serie de experimentos para comprobar el comportamiento de métodos de sobre-muestreo en contextos de *Big Data* y para conjuntos de datos pequeños, como ROS y algunas variantes de SMOTE.

Para examinar el comportamiento del clasificador, se aplicó un árbol de decisión, en un total de 12 conjuntos de datos de dos clases desbalanceados. Los resultados han demostrado que los métodos basados en SMOTE obtienen un rendimiento más bajo que ROS en escenarios de *Big Data*, mientras que en conjuntos de datos pequeños, SMOTE presenta un mejor comportamiento.

El rendimiento aparente en *Big Data* de los métodos de sobre-muestreo basados en SMOTE proviene del hecho de que esta técnica genera instancias sintéticas de acuerdo con el vecindario y la cobertura probable del espacio muestral. Los resultados sugiere que usar únicamente SMOTE para escenarios de *Big Data* no es una alternativa viable, ya que los resultados obtenidos no abordaron una mejora en el desempeño, a pesar de utilizar un método de agrupamiento para abordar la creación de datos sintéticos. Adicionalmente, se ha corroborado que el desbalance de clases no es un problema en sí mismo en los conjuntos de *Big Data*, por lo que es necesario lidiar con otras complejidades, debido a la presencia de redundancia, ruido y traslape de clases. Esta deducción se ha guiado por las métricas de complejidad F1 y F4.

Como trabajo futuro, la línea abierta se centra en considerar las regiones dispersas o las regiones densas para decidir la creación de nuevas instancias en SMOTE. Otra línea abierta involucra la reducción de características previas al uso de SMOTE, así como implementar otras métricas de distancia, como la similitud del coseno. Finalmente, la presencia de otras complejidades de datos en *Big Data* debe gestionarse antes de manejar el desbalance de clases, de este modo, buscar la generación de propuestas híbridas.

AGRADECIMIENTOS

Angélica Guzmán-Ponce contó con el apoyo del contrato postdoctoral Margarita Salas MGS/2021/23(UP2021-021) financiado por la Unión Europea-NextGenerationEU.

REFERENCIAS

Basgall, M. J., Hasperué, W., Naiouf, M., Fernández, A. & Herrera, F. (2018). SMOTE-BD: An Exact and Scalable Oversampling Method for Imbalanced Classification in Big Data. Journal of Computer Science and Technology, 18(03), 23–28. https://doi.org/10.24215/16666038.18.e23

Basgall, M. J., Hasperué, W., Naiouf, M., Fernández, A. & Herrera, F. (2019). An Analysis of Local and Global Solutions to Address Big Data Imbalanced Classification: A Case Study with SMOTE Preprocessing. Communications in Computer and Information Science, 75-85. https://doi.org/10.1007/978-3-030-27713-0_7

Batista, G. E. A. P. A., Prati, R. C. & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, 6(1), 20-29. https://doi.org/10.1145/1007730.1007735

del Río, S., López, V., Benítez, J. M. & Herrera, F. (2014). On the use of MapReduce for imbalanced big data using Random Forest. Information Sciences, 285, 112-137. https://doi.org/10.1016/j.ins.2014.03.043

García, S., Galar, M., Prati, R. C., Krawczyk, B. & Herrera, F. (2018). Learning from Imbalanced Data Sets. Springer. https://doi.org/10.1007/978-3-319-98074-4

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M. & Herrera, F. (2016). Big data preprocessing: methods and prospects. Big Data Analytics, 1(9), 1–22. https://doi.org/10.1186/s41044-016-0014-0

García, V., Sánchez, J., Marqués, A., Florencia, R. & Rivera, G. (2020). Understanding the apparent superiority of over-sampling through an analysis of local information for class-imbalanced data. Expert Systems with Applications, 158, 113026. https://doi.org/10.1016/j.eswa.2019.113026

Gutiérrez, P. D., Lastra, M., Benítez, J. M. & Herrera, F. (2017). SMOTE-GPU: Big Data preprocessing on commodity hardware for imbalanced classification. Progress in Artificial

Intelligence, 6(4), 347-354. https://doi.org/10.1007/s13748-017-0128-2

Kitchin, R. & Lauriault, T. P. (2015). Small data in the era of big data. GeoJournal, 80(4), 463-475. https://doi.org/10.1007/s10708-014-9601-7

Lichman, M. (s. f.). UCI Machine Learning Repository. Recuperado 5 de octubre de 2022, de https://archive.ics.uci.edu/ml/index.php

Luengo, J., García-Gil, D., Ramírez-Gallego, S., García, S. & Herrera, F. (2020). Big Data Preprocessing. Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-030-39105-8

Maillo, J., Ramírez, S., Triguero, I. & Herrera, F. (2017). kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data. Knowledge-Based Systems, 117, 3-15. https://doi.org/10.1016/j.knosys.2016.06.012

Maillo, J., Triguero, I. & Herrera, F. (2020). Redundancy and Complexity Metrics for Big Data Classification: Towards Smart Data. IEEE Access, 8, 87918-87928. https://doi.org/10.1109/access.2020.2991800

Maldonado, S., López, J. & Vairetti, C. (2019). An alternative SMOTE oversampling strategy for high-dimensional datasets. Applied Soft Computing, 76, 380-389. https://doi.org/10.1016/j.asoc.2018.12.024

Rendón, E., Alejo, R., Castorena, C., Isidro-Ortega, F. J. & Granda-Gutiérrez, E. E. (2020). Data Sampling Methods to Deal With the Big Data Multi-Class Imbalance Problem. Applied Sciences, 10(4), 1276. https://doi.org/10.3390/app10041276