Balanceo en la GEIH

Wilson Andrés Pinzon

Balanceo en la GEIH

En el siguiente documento se presenta un método que busca tratar el problema del desbalance respecto a la variable `ACTIVIDAD_OCUPADA_ULTIMA_SEMANA`, la cual será la variable dependiente de un modelo de regresión logística que busca explicar qué impacto tienen algunas condiciones sociales y demográficas en la probabilidad de que un joven de 18 a 28 años que pertenece a la fuerza de trabajo se encuentre trabajando.

Este método consiste en el uso de algún método de balanceo de tipo oversampling sobre la clase minoritaría, como SMOTE o ADASYN, para generar un conjunto de muestras sintéticas. Posteriormente, se utiliza el método *Propensity Score Adjusted (PSA)* para generar unos *pseudopesos* que permitan ponderar el conjunto de muestras sintéticas para realizar una estimación híbrida del modelo de regresión logística que combina los datos del conjunto original y el conjunto de muestras sintéticas.

En el siguiente documento se evaluaran tres variaciones del método SMOTE para tratar el problema de balanceo de datos, los cuales son:

1. SMOTE - ENC
2. ADASYN modificado con SMOTE-ENC
3. SMOTE-ENC ajustado con pesos de muestreo

Posteriormente, se aplicará el método de PSA para generar los pseudopesos para cada conjunto de muestras sintéticas generados, finlamente se obtendran tres estimaciones del modelo de regresión logística donde se evaluara el desempeño obtenido a través de cada método de balanceo, de esta manera se podrá determinar qué variación de SMOTE junto al método de PSA se ajusta mejor al problema de balanceo de datos en la GEIH.

Balanceo

El desbalance en los datos se refiere al caso en el que el conjunto de datos no tiene un representación equitativa de instancias que pertenecen a cada clase de la variable dependiente, en este caso, se presenta un desbalance de dos clases, en donde hay una gran diferencia en el número total de instancias que pertenecen a una clase respecto a la otra, comunmente se les denomina clase mayoritaría y minoritaria.

Actualmente el total de instancias del conjunto de datos de la GEIH que pertenecen a cada clase de la variable dependiente es la siguiente:

```
table(data$ACTIVIDAD_OCUPADA_ULTIMA_SEMANA)
```

```
##
##    0    1
## 1255 5647
```

Donde el valor $'0'$ indica que el joven está buscando trabajo mientras que el valor $'1'$ indica que ya se encuentra trabajando. Con esto en cuenta, se puede observar que hay un desbalance en los datos donde la clase minoritaría se refiere a la clase donde los jovenes que están buscando trabajo.

El tema de desbalance de los datos es un tema desafiante en el Machine Learning. Según Mukherjee et.al (2021), "los algoritmos de machine learning tienden a predecir cualquier instancia como un elemento de la clase mayoritaria, haciendo que el modelo resulte ineficiente para identificar las instancias de la clase minoritaria, esto es algo crítico, especialmente, cuando hay un gran interes en clasificar de manera correcta esta clase".

Se han desarrollado varías formas para tratar este problema, una de las formas principales consiste en realizar un re-muestreo del conjunto de datos, esto puede ser a través del oversampling de la clase minoritaría o del subsampling de la clase mayoritaría. Entre estos métodos de re-muestreo, uno de los métodos más utilizados es el método SMOTE (Chawla et al. 2002).

SMOTE es un algoritmo en el cual la clase minoritaría recibe un oversampling a través de la creación de muestras "*sintéticas*" que se ubican en los segmentos que unen a cada instancia de la clase minoritatía con sus k vecinos más cercanos en cada variable o característica.

SMOTE ha ganado una gran popularidad entre los métodos que existen para tratar el problema de balanceo y de hecho, se ha establecido como uno de los métodos más utilizados para tratar este problema. Además, desde su desarrollo han salido multiples variantes cómo los con Borderline-SMOTE, ADASYN, SMOTE ENN, entre otros, que buscan mejorar su rendimiento endiferentes escenarios.

Ahora, un problema que tienen estos métodos, es que fueron desarrollados bajo la consideración que todas las variables son continua, en el caso que se trabaje sobre un conjunto de datos que contiene variables nominales, como es el caso de la GEIH, tanto SMOTE como sus variantes no son directamente aplicables. Si bien existe una variante en donde se codifican las variables nominales a través de la técnica One Hot Encoding, esta variante no es la mejor solución ya que aumenta considerablemente el costo computacional del algoritmo y además, es posible que el algoritmo no aprenda sobre las posibles relaciones entre los valores nominales y las clases.

Por este motivo, se han desarrollado variantes de SMOTE que permiten manejar variables nominales y continuas. Una de estas variantes es SMOTE-ENC(SMOTE Encoded Nominal and Continuous) (Mukherjee and Khushi 2021), en donde las variables nominales son codificadas como variables numéricas y en donde un valor más alto representa una asociación más fuerte con la clase minoritaría.

Este algoritmo será el primer método que se utilizará como método para tratar el problema de balanceo en la GEIH, para esto se generará el conjunto de muestras sintéticas sobre un conjunto de entrenamiento:

```
### Método SMOTE ENC
synt.smote = SMOTE_ENC(train.nowt, target, minority.value, vars.numeric, k , seed)
```

SMOTE-ENC es una alternativa que nos permite tratar el problema del desbalance en conjuntos con variables numéricas y nominales, sin embargo, es cabe resaltar que este método proviene de SMOTE y por lo tanto, puede heredar algunas de las limitaciones de este método.

En ese sentido, ADASYN (Adaptive Synthetic) (He et al. 2008) surge como una de las variantes de SMOTE más robusta. Este método se basa en la idea de generar muestras sintéticas de la clase minoritaría de forma adaptativa, es decir, busca generar más muestras sintéticas de auqellas intancias con una menor densidad.

La mayor diferencia entre SMOTE y ADASYN radica en que SMOTE genera la misma cantidad de registros sintéticos para cada muestra de la clase minoritaria, mientras que ADASYN provee un peso a cada registro de la clase minoritaria para determinar el número de muestras sintéticas que deben ser generadas por cada registro.

Ahora, como se comento anteriormente, ciertas variantes de SMOTE, incluyendo ADASYN, funcionan bajo la consideración que todas las variables son numéricas, así que, para tratar este problema, se propone realizar

una variante del algoritmo ADASYN, donde se planea utilizar la métrica dispuesta en el método SMOTE-ENC para encontrar los k vecinos más cercanos de cada instancia de la clase minoritaria. La utilización de esta métrica permite combinar la robustez de ADASYN con la capacidad que tiene SMOTE-ENC para tratar variables numéricas y nominales.

Esta variante de ADASYN con SMOTE-ENC será el segundo método utilizado dentro del documento para tratar el problema de desbalance de datos en la GEIH, se generará el conjunto de muestras sintéticas sobre el mismo conjunto de entrenamiento utilizado en SMOTE-ENC.

```
### Método ADASYN con SMOTE-ENC
synt.adasyn = ADASYN(train.nowt, target, minority.value, vars.numeric, k , seed)
```

Además de los métodos SMOTE-ENC y ADASYN, este documento propone una tercera alternativa para tratar el problema del desbalance de datos, esta alternativa está basado en el algoritmo WSMOTE (Prusty, Jayanthi, and Velusamy 2017).

WSMOTE es un método de sobre muestreo que asigna unos pesos a cada instancia y que determinan el número de muestras sintéticas que se van a genrar a través de SMOTE para cada instancia de la clase minoritaria.

La variante propuesta consiste en adaptar la idea del uso de unos pesos para determinar el número de muestras sintéticas, sin embargo, en el caso propuesto, se elimina el calculo de los pesos explícitos, en su lugar, se van a utilizar los pesos de muestreo asociados al conjunto de datos para determinar el número de muestras sintéticas que se van a generar de cada instancia. Se propone además, para poder tener un número adecuado de muestras sintéticas, normalizar los pesos de muestreo.

Esta propuesta se basa en la idea de que los pesos de muestreo representan el porcentaje de la población que representa cada registro dentro del conjunto de datos. Utilizar estos pesos normalizados, permite generar un conjunto de muestras sintéticas que procura preservar las distribuciones asociadas a la población original.

Otra consideración es que WSMOTE es una variante que determina la cantidad de muestras sintéticas que se van a generar de cada instancia, pero, genera estas muestras a través de SMOTE, por lo cual, trabaja bajo la consideración que todas las variables son numéricas, en ese sentido, la propuesta de SMOTE con Pesos de muestreo también utilizará la metrica de SMOTE-ENC para encontrar los k vecinos más cercanos.

La generación del conjunto de muestras sintéticas para esta propuesta de método también será realizada a partir del conjunto de entrenamiento:

```
factor.expansion <- train$FACTOR_EXPANSION
synt.swsmoteenc = SWSMOTEENC(train.nowt, factor.expansion, target, minority.value, vars.numeric, k , se
```

Ahora, un detalle muy importante a tener en cuenta es que el conjunto de datos de la GEIH, es un conjunto de datos basado en un muestreo complejo, por lo que el peso de muestreo tiene un papel fundamental sobre cualquier tipo de análisis que se desee realizar. Este valor permite ajustar los resultados del conjunto de datos para obtener estimaciones más precisas sobre la población de estudio, y de hecho, ignorar esta variable puede llevar a realizar estimaciones imprecisas.

En el contexto del desbalance de datos, este peso de muestreo es un factor fundamental. SMOTE muestra probabilística y no probabilística.

Para resolver este problema, a través del Propensity Score (Ebrahim Valojerdi and Janani 2018)

Evaluación de los modelos

Referencias

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic Minority over-Sampling Technique." *Journal of Artificial Intelligence Research* 16 (June): 321–57. https://doi.

org/10.1613/jair.953.

Ebrahim Valojerdi, Ameneh, and Leila Janani. 2018. "A Brief Guide to Propensity Score Analysis." *Medical Journal of the Islamic Republic of Iran*, September, 717–20. https://doi.org/10.14196/mjiri.32.122.

He, Haibo, Yang Bai, Edwardo A. Garcia, and Shutao Li. 2008. "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning." In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE. https://doi.org/10.1109/ijcnn.2008.4633969.

Mukherjee, Mimi, and Matloob Khushi. 2021. "SMOTE-ENC: A Novel SMOTE-Based Method to Generate Synthetic Data for Nominal and Continuous Features." *Applied System Innovation* 4 (1): 18. https://doi.org/10.3390/asi4010018.

Prusty, Manas Ranjan, T. Jayanthi, and K. Velusamy. 2017. "Weighted-SMOTE: A Modification to SMOTE for Event Classification in Sodium Cooled Fast Reactors." *Progress in Nuclear Energy* 100 (September): 355–64. https://doi.org/10.1016/j.pnucene.2017.07.015.