ISSN 1870-4069

Rendimiento de los métodos de sobremuestreo para los modelos predictivos del Síndrome de Guillain-Barré

Manuel Torres-Vásquez[1,2], José Hernández-Torruco[1], Betania Hernández-Ocaña[1], Oscar Chávez-Bosquez[1]

[1] Universidad Juárez Autónoma de Tabasco, División Académica de Ciencias y Tecnologías de la Información, México
[2] Tecnológico Nacional de México/ITS de Centla, División Sistemas Computacionales, México

{jose.hernandezt, betania.hernandez, oscar.chavez}@ujat.mx, mtorresv@itscentla.edu.mx

Resumen. El Síndrome de Guillain-Barré (SGB) es un trastorno neurológico donde el sistema inmune del cuerpo ataca al sistema nervioso periférico. Esta enfermedad es de rápida evolución y puede llevar a una parálisis de todo el cuerpo. En casos graves necesita ventilación asistida. Existen cuatro variantes del SGB: Polineuropatía Desmielinizante Inflamatoria Aguda (AIDP), Neuropatía Axonal Motora Aguda (AMAN), Neuropatía Axonal Sensorial Aguda (AMSAN) y Síndrome de Miller-Fisher (MF). Es necesario identificar el subtipo de SGB que el paciente contrajo debido a que el tratamiento es diferente de acuerdo al subtipo contraído. Esta enfermedad puede afectar la vida y la familia del paciente debido a que puede llevar años la recuperación total. Realizamos este estudio experimental con un *dataset* desbalanceado multiclase. Este *dataset* es un registro de 129 pacientes diagnosticados con alguna variante del SGB. Binarizamos el *dataset* original aplicando la técnica "Uno contra todos". Aplicamos tres métodos de sobremuestreo (MWMOTE, RWO y RACOG), los cuales generan instancias sintéticas usando diferente enfoque. Aplicamos tres clasificadores individuales (Árbol de decisión, SVM y JRip). El rendimiento de los algoritmos se obtuvo mediante la curva ROC y el área bajo la curva AUC. Se obtuvieron modelos predictivos con datos balanceados y modelos predictivos con datos desbalanceados. Aplicamos la prueba Wilcoxon para conocer si existen diferencias estadísticamente significativas entre los modelos balanceados y los modelos desbalanceados. Los resultados muestran que balancear el *dataset* mejora el rendimiento de los modelos predictivos. Balancear la clase minoritaria utilizando el método RWO en combinación con el clasificador SVM obtuvo los mejores resultados.

Palabras clave: Dataset multiclase, desbalanceo, sobremuestreo, aprendizaje automático, clasificación, binarización.

Manuel Torres-Vásquez, José Hernández-Torruco, Betania Hernández-Ocaña, Oscar Chávez-Bosquez

Performance of Oversampling Methods for Guillain-Barré Syndrome Predictive Models

Abstract. Guillain-Barré Syndrome (GBS) is a neurological disorder where the body's immune system attacks the peripheral nervous system. This disease is rapidly evolving and can lead to paralysis of the entire body. In severe cases, it needs assisted ventilation. There are four GBS variants: Acute Inflammatory Demyelinating Polyneuropathy (AIDP), Acute Motor Axonal Neuropathy (AMAN), Acute Sensory Axonal Neuropathy (AMSAN), and Miller-Fisher Syndrome (MF). It is necessary to identify the GBS subtype that the patient has because the treatment is different according to the developed subtype. This disease can affect the life and family of the patient because it can take years for the patient to fully recover. We conducted this experimental study with an imbalanced multiclass dataset. This dataset is a record of 129 patients diagnosed with some variant of GBS. We converted in binary the original dataset applied the technique "One versus all". We applied three oversampling methods (MWMOTE, RWO, and RACOG) which generate synthetic instances with a different approach. We applied three simple classifiers (decision tree, SVM and JRip). The performance of the algorithms was obtained using ROC curve and the area under the curve AUC. Predictive models with balanced data were obtained, along with predictive models with imbalanced data. We applied the Wilcoxon test to find out if there were statistically significant differences between the balanced models and the imbalanced ones. Results show that balancing the dataset improves the performance of predictive models. Balancing the minority class using the RWO method in combination with the SVM classifier obtained the best results.

Keywords: Multiclass dataset, imbalance, oversampling, machine learning, classification, binarization.

1. Introducción

El Síndrome de Guillain-Barré es una rara afección que ocurre cuando el sistema de defensa del cuerpo, comúnmente llamado sistema inmunitario, ataca parte del sistema nervioso periférico. Los primeros síntomas son debilidad muscular y hormigueo en las extremidades. La rápida evolución de este padecimiento puede llevar a paralizar todo el cuerpo. Existen cuatro variantes conocidas: Polineuropatía Desmielinizante Inflamatoria Aguda (AIDP), Neuropatía Axonal Motora Aguda (AMAN), Neuropatía Axonal Sensorial Aguda (AMSAN) y Síndrome de Miller-Fisher (MF). Se desconoce la causa exacta del SGB, pero del 50 % al 70 % de los casos aparecen de una a dos semanas después de una infección respiratoria o gastrointestinal [7]. Actualmente existen estudios en los cuales se relaciona con el Zika [15].

Los casos graves son raros, sin embargo, pueden producir una parálisis casi total y requerir una monitorización de cuidados intensivos prolongada e incluso ventilación mecánica, aunque la mayoría de los casos se recuperan totalmente.

Esta enfermedad puede afectar extremadamente la vida y la familia de un paciente porque puede llevar años la recuperación total del individuo [19]. El ingreso temprano al hospital y el tratamiento temprano son importantes para disminuir la necesidad de asistencia respiratoria y mejorar el resultado. El diagnóstico se basa en hallazgos clínicos, de laboratorio y neurofisiológicos. El tratamiento temprano es crucial para un resultado favorable. El tratamiento varía de acuerdo al subtipo contraído. La identificación del subtipo de SGB que el paciente contrajo en forma temprana, permitirá aplicar el tratamiento adecuado para disminuir la morbilidad y mortalidad a largo plazo [10].

En trabajos anteriores [2,3] se crearon modelos predictivos utilizando los datos originales, es decir, un conjunto de datos conformado por clases desbalanceadas. El desbalanceo de clases es un problema que afecta el resultado de los clasificadores. Este se presenta debido a que los algoritmos de clasificación están basados en la idea que los *datasets* se encuentran balanceados. Cuando un clasificador utiliza datos desbalanceados tiende a ignorar la clase minoritaria y sesgar su resultado en la clase mayoritaria. El desbalanceo de datos se presenta cuando en un *dataset* la cantidad de datos que conforma una de las clases tiene significativamente menos datos respecto a la otra clase [11].

Existen dos tipos de desbalanceo de datos. El primero se conoce como desbalanceo binario. El desbalanceo binario está formado por un *dataset* con dos clases, sin embargo, una de ellas contiene un mayor número de datos con respecto a la otra. A esta clase se le llama clase mayoritaria y a la clase que tiene un menor número de instancias se le conoce como clase minoritaria. El segundo tipo de desbalanceo, es el desbalanceo multiclase. El desbalanceo multiclase se presenta cuando el *dataset* está formado por más de dos clases y su distribución de datos es desigual para cada una de las clases. Para solventar el problema del desbalanceo de datos, han surgido varias técnicas que mejoran el rendimiento de los clasificadores.

Técnica a nivel algoritmo: esta técnica modifica el algoritmo de clasificación para darle mayor peso a la clase minoritaria para ser más preciso. Sin embargo, esta técnica no es variable, debido que depende de un clasificador específico y las modificaciones están destinadas a resolver un problema de desbalanceo para un algoritmo en particular [11].

Técnica a nivel de datos: basándose en el hecho que los datos son los que deben estar equilibrados, este método realiza a través de diferentes técnicas el balanceo de datos antes de pasarlos por el clasificador. A este método se le llama preprocesamiento de datos. La ventaja de esta técnica es que puede ser utilizada prácticamente con cualquier clasificador [8].

Técnica costo-sensitivo: este método asigna pesos o instancias al conjunto de prueba y no al conjunto de entrenamiento como en el modelo a nivel de datos [8].

En la técnica a nivel de datos existen tres diferentes formas de balancear los datos de entrenamiento. El submuestreo de clases, que elimina instancias de la clase mayoritaria hasta equilibrarla con la clase minoritaria. Sin embargo, no es recomendable para un *dataset* con pocos datos debido a la pérdida de datos. El sobremuestreo es una de las técnicas más utilizadas para balancear los datos. Esta técnica genera nuevas instancias para anexarlas a la clase minoritaria y equilibrarla con la clase mayoritaria. Por último, existen métodos híbridos que combinan el sobremuestreo y el submuestreo. En la literatura especializada existen diferentes métodos de sobremuestreo clásicos para balancear la clase minoritaria del conjunto de entrenamiento.

Chawla y colaboradores [4] presentan SMOTE *(Synthetic Minority Oversampling Technique)*, este método es el más usado y con mayor éxito para balancear un *dataset*. SMOTE crea instancias sintéticas a lo largo de los segmentos de la línea que unen a cualquiera de los vecinos más cercanos a la clase k minoritaria. Las nuevas instancias se generan a través de diferencias entre las instancias y su vector de característica considerando sus vecinos más cercanos. Existen métodos basados en SMOTE creados para mejorar la generación de instancias sintéticas, pero SMOTE es el considerado clásico en la literatura.

En un primer estudio exploratorio [17], analizamos el comportamiento de los datos balanceados aplicando el método SMOTE con seis diferentes porcentajes de sobremuestreo sintético para la clase minoritaria, binarizando el dataset aplicando OVO. Los resultados arrojaron que SMOTE al 100 % obtuvieron los mejores resultados. En un estudio más amplio [18], utilizamos dos técnicas diferentes de binarización: OVO (Uno contra Uno) y OVA (Uno contra todos). Los resultados demostraron que eliminar datos baja el rendimiento de los clasificadores, y SMOTE volvió a demostrar su eficacia para mejorar el rendimiento de los clasificadores respecto a los datos desbalanceados.

En este trabajo cumplimos dos objetivos. El primero fue explorar tres diferentes métodos actuales de sobremuestreo para conocer cuál de ellos mejora el rendimiento de los clasificadores, y saber si estos métodos superan el rendimiento de los métodos utilizados en trabajos anteriores. Estos métodos crean instancias sintéticas de diferente manera. El segundo, conocer si crear modelos predictivos con *datasets* balanceados mejoran los modelos predictivos creados con *datasets* desbalanceados realizados en trabajos anteriores.

Para esto, convertimos el *dataset* multiclase desbalanceado en *subsets* binarios utilizando la técnica de binarización "Uno contra todos". Para realizar el balanceo de datos, aplicamos tres métodos de sobremuestreo: MWMOTE [1], RWO [21] y RACOG [6], que generan instancias sintéticas con diferente enfoque cada uno. Las instancias creadas se anexan a la clase minoritaria hasta alcanzar el equilibrio con la clase mayoritaria. Aplicamos tres algoritmos de clasificación que crean modelos predictivos de diferente manera. C4.5 (Árbol de decisión), forma jerárquica; SVM basado en kernel; JRip *(Ripper)* basado en reglas. Utilizamos la curva ROC y el área bajo la curva AUC como medidas de rendimiento. Los modelos creados son el promedio de 60 ejecuciones con diferentes semillas.

Se aplicó la prueba Wilcoxon, para conocer si existe una diferencia estadísticamente significativa entre los modelos predictivos creados con datos balanceados versus los modelos predictivos creados con los datos desbalanceados.

Este artículo está organizado de la siguiente forma: En la sección 2 presentamos una descripción del *dataset*, los algoritmos de clasificación y los métodos de sobremuestreo y la medida de rendimiento utilizada en el estudio. La sección 3 describe el procedimiento experimental. En la sección 4 presentamos los resultados experimentales. Finalmente, en la sección 5 damos conclusiones del estudio y sugerimos trabajos futuros.

2. Materiales y métodos

2.1. Conjunto de datos

El *dataset* usado en este estudio es un registro de pacientes diagnosticados con SGB. Estos datos son una recopilación de información de 129 individuos con algún subtipo de SGB, divididos en: 20 con AIDP, 37 con AMAN, 59 con AMSAN y 13 con MF. Los datos originales están conformados por 4 clases desbalanceadas, es decir, la cantidad de pacientes con algún subtipo de SGB es desigual respecto a las otras. El *dataset* se obtuvo a través del Instituto Nacional de Neurología y Neurocirugía de la Ciudad de México.

El *dataset* original consta de 356 variables. En un artículo anterior seleccionamos las variables más relevantes [9]. Las variables v22, v29, v30 y v31 son de tipo clínico. El resto de variables pertenecen a la prueba de conducción nerviosa. A continuación, se muestran las variables utilizadas en este estudio.
`[v22]`= Simetría (en debilidad), `[v29]`= Afectación de los músculos extraoculares, `[v30]`= Ptosis, `[v31]`= Implicación cerebelosa, `[v63]`= Amplitud del nervio motor mediano izquierdo, `[v106]`= Área bajo la curva del nervio motor cubital izquierdo, `[v120]`= Área bajo la curva del nervio motor cubital derecho, `[v130]`= Amplitud del nervio motor tibial izquierdo, `[v141]`= Amplitud del nervio motor tibial derecho, `[v161]`= Área bajo la curva del nervio motor peroneo derecho, `[v172]`= Amplitud del nervio sensorial mediano izquierdo, `[v177]`= Amplitud del nervio sensorial mediano derecho, `[v178]`= Área bajo la curva del nervio sensorial mediano derecho, `[v186]`= Latencia del nervio sensorial cubital derecho, `[v187]`= Amplitud del nervio sensorial cubital derecho, `[v198]`= Área bajo la curva del nervio sensorial sural derecho.

2.2. Algoritmos de aprendizaje automático

Algoritmos de sobremuestreo

- MWMOTE *(Majority Weighted Minority Oversampling TEchnique)*: Está basado en que no solo el desbalanceo de clases afecta el resultado del clasificador, sino que también la posición de las instancias de la clase minorita repercute en el resultado. Además, en que los métodos de sobremuestreo actuales pueden producir instancias sintéticas incorrectas. El objetivo de

este método es mejorar la selección de las instancias minoritarias y mejorar la forma de generar los nuevos datos sintéticos. MWMOTEestá divido en tres etapas: 1) identifica las muestras de la clase minoritaria más importantes y difíciles de aprender del *dataset* minoritario original, 2) asigna un peso de selección de acuerdo a la importancia en los datos, 3) genera las muestras sintéticas utilizando el enfoque de agrupamiento para asegurarse que se encuentran dentro de un grupo de la clase minoritaria.
- RWO *(Random Walk Over-Sampling approach):* Utiliza un enfoque de sobremuestreo de caminata aleatoria. El objetivo es agregar nuevas instancias sintéticas lo más cerca posible de las originales. Además, respetar la distribución original de la clase minoritaria más expandir los límites de la clase minoritaria lograrán obtener un conjunto de datos balanceado más beneficioso que ayudará en el rendimiento de los clasificadores. RWO intenta mantener la distribución de los datos de la clase minoritaria sin cambios calculando la media y la varianza de los atributos numéricos.
- RACOG *(RApidy COnverging Gibbs):* Utiliza la distribución de probabilidad de la clase minoritaria para seleccionar y crear estratégicamente nuevas muestras de entrenamiento de clase minoritaria. RACOG utiliza un esquema de muestreo de Gibbs para generar una cadena de Márkov con nuevas muestras sintéticas de clase minoritaria utilizando el algoritmo del árbol de dependencia de Chow-Liu. Este método evita la superposición de clases.

Utilizamos estos métodos de sobremuestreo para balancear los datos de entrenamiento debido a que son algoritmos creados recientemente. Además, cada uno de ellos genera de forma diferente instancias sintéticas para la clase minoritaria.

Clasificadores individuales

- C4.5 (Árbol de decisión): Este clasificador basa su resultado en forma jerárquica. Es fácil de interpretar, rápido y robusto al ruido. Es recomendado para diagnóstico médico [8].
- SVM (Máquina de vector soporte): SVM construye un hiperplano en un espacio bidimensional para separar las clases. Es uno de los clasificadores con muy alta precisión respecto a otros clasificadores, como regresión logística y Naive Bayes [8].
- JRip (RIPPER, *repeated incremental pruning to produce error reduction*): Identifica un conjunto de reglas para obtener el resultado. Es fácil de interpretar para los humanos [16].

Seleccionamos estos algoritmos con el objetivo de obtener los resultados usando diferentes enfoques de clasificación.

2.3. Medidas de rendimiento

Evaluamos la eficiencia de los modelos predictivos mediante la curva ROC *(Receiver Operating Characteristic)* y el área bajo la curva AUC *(Area Under*

the Curve). La curva ROC mide qué tan bien se clasifican las predicciones, así como la calidad de las predicciones del modelo. La curva ROC se define como la sensibilidad, que es la tasa de verdadero positivos y 1-especificidad que es la tasa de falsos positivos. El área bajo la curva AUC nos permite identificar una clase. Para este experimento, nos sirve para discriminar entre un subtipo de SGB entre los otros subtipos. En esta medida de rendimiento, los valores $\geq .900$ se consideran modelos excelentes. Los valores $\geq .700$ significa que son buenos modelos. Sin embargo, los valores con $\leq .500$ están considerados malos modelos [14].

2.4. Técnica de binarización

En problemas de clasificación con *datasets* multiclase es común convertir el *dataset* original en subproblemas binarios. Una de las formas generalmente utilizadas es la técnica "Uno contra todos" [13]. Esta técnica toma una clase que se convertirá en la clase minoritaria. La suma de las clases restantes serán la clase mayoritaria. Este proceso se repetirá dependiendo el número de clases que tenga el *dataset* original.

Utilizamos la técnica de binarización "Uno contra todos" para reconstruir el *dataset* original en varios *datasets* binarios:

El primer *subset* está formado por las instancias de la clase 1 (AIDP), llamada clase minoritaria. La clase mayoritaria se forma a través de la suma de las instancias de las clases restantes 2 (AMAN), 3 (AMSAN) y 4 (MF).

El segundo *subset* está formado por las instancias de la clase 2 (AMAN), llamada clase minoritaria. La clase mayoritaria se forma a través de la suma de las instancias de las clases restantes 1 (AIDP), 3 (AMSAN) y 4 (MF).

El tercer *subset* está formado por las instancias de la clase 3 (AMSAN), llamada clase minoritaria. La clase mayoritaria se forma a través de la suma de las instancias de las clases restantes 1 (AIDP), 2 (AMAN) y 4 (MF).

El cuarto *subset* está formado por las instancias de la clase 4 (MF), llamada clase minoritaria. La clase mayoritaria se forma a través de la suma de las instancias de las clases restantes 1 (AIDP), 2 (AMAN) y 3 (AMSAN).

Utilizar esta técnica nos ayuda a identificar uno de los cuatro subtipos de SGB, discriminando de los restantes.

2.5. Validación

Utilizamos la evaluación *train-test* para cada clasificador. Dividimos el *dataset* en dos subconjuntos de datos. El primero son los datos de entrenamiento, estos se usaron para construir el modelo. El segundo conjunto de datos llamado prueba, se mantienen aparte y a través de ellos se evaluó el modelo. El rendimiento de este modelo se realizó con los datos de prueba. Usamos dos tercios de los datos para el entrenamiento y un tercio para las pruebas del modelo.

3. Procedimiento experimental

Utilizamos la técnica de binarización "Uno contra todos" para crear cuatro *subsets* binarios desbalanceados. En la Tabla 1 se muestra el resultado de la binarización del *dataset* original multiclase desbalanceado SGB. La primera columna, muestra los cuatro *subsets* obtenidos con la técnica "Uno contra todos". La segunda columna, muestra el número de instancias que forman la clase minoritaria para cada *subset*. La tercera columna, muestra el subtipo que pertenece la clase minoritaria. La cuarta columna, muestra el número de instancias que forman la clase mayoritaria para cada *subset*. La quinta columna, muestra los subtipos que forman la clase mayoritaria.

Tabla 1. *Subsets* binarios obtenidos con el método "Uno contra todos".

Subset	**Clase minoritaria**	**Subtipo**	**Clase mayoritaria**	**Suma subtipos**
$SGB1$	20	AIDP	109	AMAN, AMSAN y MF
$SGB2$	37	AMAN	92	AIDP, AMSAN y MF
$SGB3$	59	AMSAN	70	AIDP, AMAN y MF
$SGB4$	13	MF	116	AIDP, AMAN y MF

Tabla 2. Resultados de los *subsets* balanceados aplicando los métodos de sobremuestreo.

Subset	**Datos de entrenamiento desbalanceados**	**Datos generados con MWMOTE, RWO y RACOG**	**Clase minoritaria balanceada**	**Clase mayoritaria original**
$SGB1$	15	56	71	71
$SGB2$	25	37	62	62
$SGB3$	40	7	47	47
$SGB4$	9	69	78	78

Particionamos cada *subset* en entrenamiento (66 %) y prueba (34 %). A los cuatro *subsets* de entrenamiento de la clase minoritaria, se le aplicaron tres diferentes métodos de sobremuestreo (MWMOTE, RWO y RACOG) con el objetivo de equilibrarlos con la clase mayoritaria. Los datos de prueba se utilizaron para medir el rendimiento de los modelos obtenidos. En la Tabla 2 se muestran los cuatro *subsets* balanceados con las tres diferentes técnicas de sobremuestreo. En la primera columna se muestran los cuatro *subsets*. La segunda columna, muestran los datos de entrenamiento de la clase minoritaria. La tercera columna, muestra las instancias generadas a través de cada método de sobremuestreo.

La cuarta columna, muestra la suma de las instancias originales con las instancias generadas por los métodos de sobremuestreo. La quinta columna, muestra los datos de entrenamiento de la clase mayoritaria.

Una vez balanceados los cuatro *subsets* se les aplicaron los tres clasificadores individuales (C4.5, SVM y JRip). Cada uno de los modelos es el promedio de 60 ejecuciones con diferente semilla. La medida de rendimiento fue el área bajo la curva AUC. Por otro lado, se utilizaron los *subsets* desbalanceados para aplicar los tres clasificadores individuales. Se obtuvieron modelos predictivos con los datos desbalanceados.

Realizamos una comparación de los modelos obtenidos con los datos balanceados contra los modelos conseguidos con los datos desbalanceados. Se aplicó la prueba Wilcoxon para conocer si existe diferencia estadísticamente significativa, siempre y cuando los modelos balanceados superaran los modelos desbalanceados.

Los experimentos en R se realizaron sobre RStudio 1.2.1335. Utilizamos el paquete `imbalance` [5] para los tres métodos de sobremuestreo MWMOTE, RACOG y RWO. Respecto a los clasificadores, utilizamos el paquete `RWeka 0.4-39` [20] para C4.5 y JRip. El clasificador SVM usamos el paquete `e1071 1.7-0` [12].

SVM se optimizó a través de la función *tune*, asignando los valores 0.001, 0.01, .1, 1, 10, 50, 80, 100 para el parámetro C.

4. Resultados y discusión

La Tabla 3 muestra los modelos predictivos obtenidos mediante la aplicación de tres diferentes métodos de sobremuestreo (MWMOTE, RWO y RACOG). Utilizamos la curva ROC y el área bajo lacurva AUC como métrica de los modelos. Cada modelo es el promedio de los resultados obtenidos a través de 60 ejecuciones. Se realizó la prueba estadística Wilcoxon de los modelos balanceados contra los modelos desbalanceados, siempre y cuando el modelo balanceado haya obtenido un mejor rendimiento respecto al modelo desbalanceado.

El objetivo fue conocer si existe una diferencia estadísticamente significativa entre dichos modelos. En la columna 1 se muestran los cuatro *subsets* obtenidos mediante la binarización del *dataset*. La columna 2 muestra los tres clasificadores utilizados en cada *subset* para obtener los modelos predictivos. En la columna 3 se muestran los modelos obtenidos con los clasificadores utilizando datos desbalanceados. En las columnas 4, 6 y 8 se observan los modelos obtenidos usando datos balanceados aplicando tres diferentes técnicas de sobremuestreo y tres clasificadores. Las columnas 5, 7 y 9 muestra el resultado de la prueba Wilcoxon. NPW representa que no fue necesario realizar la prueba Wilcoxon; $\mathcal{NS}$ significa que no se encontró diferencia significativa entre los modelos balanceados y los modelos desbalanceados; $\boldsymbol{\mathcal{S}}$ muestra que existe diferencia estadísticamente significativa entre los modelos balanceados y los modelos desbalanceados. Se obtuvieron 36 modelos balanceados creados con los tres diferentes métodos de sobremuestreo (MWMOTE, RWO y RACOG) y tres clasificadores (C4.5, SVM y

JRip). Por otro lado, se crearon 12 modelos utilizando los datos desbalanceados. Se realizó una comparación entre los modelos desbalanceados y los modelos balanceados. La comparación se realizó a través de la prueba estadística Wilcoxon para conocer si existen diferencias estadísticamente significativas entre dichos modelos. Con cada *subset* se crearon 9 modelos balanceados y 3 modelos desbalanceados

Usando el *subset* SGB1 los resultados muestran que el método MWMOTE con el clasificador JRip obtuvieron un modelo con diferencia estadísticamente significativa. El clasificador SVM con los datos balanceados utilizando los métodos MWMOTE y RWO, obtuvieron dos modelos que superaron el rendimiento de los datos desbalanceados, sin embargo, no se encontraron diferencias estadísticamente significativas. En los 6 modelos restantes no se realizaron prueba Wilcoxon debido que los datos desbalanceados tuvieron un mejor rendimiento respecto a los datos balanceados.

Con el *subset* SGB2 los resultados muestran que el método RWO utilizando el clasificador SVM obtuvo una diferencia estadísticamente significativa. El clasificador SVM en combinación con RACOG mejoró los datos desbalanceados; el clasificador JRip aplicando MWMOTE y RWO obtuvieron dos modelos que mejoraron el rendimiento de los modelos desbalanceados; El clasificador SVM usando RACOG obtuvo un modelo que mejoró los modelos desbalanceados, sin embargo, en ninguno de los casos se encontraron diferencias estadísticamente significativas. En otros cinco modelos balanceados no se realizó la prueba Wilcoxon debido que los modelos con datos desbalanceados obtuvieron un mejor rendimiento respecto a los modelos balanceados.

Aplicando el *subset* SGB3 los resultaron muestran que 6 modelos balanceados superaron a los modelos desbalanceados. El método de sobremuestreo RACOG con los tres clasificadores, los tres modelos balanceados mejoraron a los modelos desbalanceados. El método RWO con los clasificadores SVM y JRip, dos modelos balanceados superaron a los modelos desbalanceados. El método MWMOTE utilizando el clasificador SVM, un modelo balanceado mejoró los modelos desbalanceados. Sin embargo, en los seis casos no se encontraron diferencias estadísticamente significativas. En tres modelos balanceados no se realizó la prueba Wilcoxon debido que los modelos con datos desbalanceados obtuvieron un mejor rendimiento respecto a los modelos balanceados. Este *subset* obtuvo el peor rendimiento. En ninguno de los 9 modelos creados con datos balanceados se encontró diferencia estadísticamente significativa.

Finalmente, con el *subset* SGB4, en todos los casos los modelos balanceados superaron el rendimiento de los modelos desbalanceados. El método RWO con los tres clasificadores,tres modelos obtuvieron diferencias estadísticamente significativas. MWMOTE en combinación con C4.5 y SVM, dos modelos obtuvieron diferencias estadísticamente significativas. RACOG usando SVM obtuvo un modelo con diferencia estadísticamente significativa. Los tres modelos balanceados restantes mejoraron el rendimiento de los datos desbalanceados, sin embargo, no se obtuvieron diferencias estadísticamente significativas. Este *subset* obtuvo el mejor rendimiento respecto a los demás.

Tabla 3. Resultados de los modelos predictivos utilizando MWMOTE, RWO y RACOG para sobremuestrar la clase minoritaria. Los valores son el promedio del área bajo la curvas AUC obtenidas (NPW = No fue necesaria la prueba de Wilcoxon; $\mathcal{NS}$ = No significantivo; $\boldsymbol{\mathcal{S}}$ = Significantivo).

Sub-conjuntos	**Clasificador**	**Datos desbalan-ceados**	**Balanceo aplicando MWMOTE**	**Resultado Prueba Wilcoxon**	**Balanceo aplicando RWO**	**Resultado Prueba Wilcoxon**	**Balanceo aplicando RACOG**	**Resultado Prueba Wilcoxon**
SGB1	C4.5	0.8130	0.8042	NPW	0.7826	NPW	0.7838	NPW
	SVM	0.7477	0.7542	$\mathcal{NS}$	0.7681	$\mathcal{NS}$	0.7380	NPW
	JRip	0.7826	**0.8275**	$\boldsymbol{\mathcal{S}}$	0.7639	NPW	0.7745	NPW
SGB2	C4.5	0.9003	0.8957	NPW	0.8954	NPW	0.8749	NPW
	SVM	0.8594	0.8593	NPW	**0.9082**	$\boldsymbol{\mathcal{S}}$	0.8625	$\mathcal{NS}$
	JRip	0.8608	0.8667	$\mathcal{NS}$	0.8826	$\mathcal{NS}$	0.8444	NPW
SGB3	C4.5	0.8632	0.8629	NPW	0.8495	NPW	0.8683	$\mathcal{NS}$
	SVM	0.7898	0.7901	$\mathcal{NS}$	0.7960	$\mathcal{NS}$	0.7903	$\mathcal{NS}$
	JRip	0.8470	0.8461	NPW	0.8494	$\mathcal{NS}$	0.8479	$\mathcal{NS}$
SGB4	C4.5	0.7662	**0.8271**	$\boldsymbol{\mathcal{S}}$	**0.8639**	$\boldsymbol{\mathcal{S}}$	0.7746	$\mathcal{NS}$
	SVM	0.6846	**0.7917**	$\boldsymbol{\mathcal{S}}$	**0.8787**	$\boldsymbol{\mathcal{S}}$	**0.7729**	$\boldsymbol{\mathcal{S}}$
	JRip	0.8319	0.8509	$\mathcal{NS}$	**0.8845**	$\boldsymbol{\mathcal{S}}$	0.8537	$\mathcal{NS}$

Tabla 4. Resumen de los modelos con diferencia estadísticamente significativa, por método y por clasificador.

Subsets	**MWMOTE**	**RWO**	**RACOG**
$SGB1$	JRip	-	-
$SGB2$	-	SVM	-
$SGB3$	-	-	-
$SGB4$	C4.5, SVM	C4.5, SVM, JRip	SVM

En la Tabla 4 se muestran los resultados de los modelos que obtuvieron diferencias estadísticamente significativas. El método RWO obtuvo el mejor rendimiento respecto a los otros dos métodos de sobremuestreo. RWO obtuvo cuatro modelos con diferencia estadísticamente significativa.

MWMOTE obtuvo 3 modelos con diferencia estadísticamente significativa. RACOG fue el método con el peor rendimiento, solo se encontró un modelo con diferencia estadísticamente significativa. En cuanto a los clasificadores, empleando SVM se hallaron 4 modelos con diferencia estadísticamente significativa. Aplicando JRip, se encontraron 2 modelos con diferencia estadísticamente significativa. El clasificador C4.5, encontró 2 modelos con diferencia estadísticamente significativa.

De los 36 modelos obtenidos con datos balanceados, 14 modelos superaron los modelos desbalanceados, sin embargo, no se encontró diferencia estadísticamente significativa. Por otro lado, 14 modelos fueron superados por los modelos desbalanceados, por lo tanto, no se aplicó la prueba Wilcoxon. Por último 8 modelos tuvieron diferencia estadísticamente significativa.

El método de sobremuestreo RWO en combinación con el clasificador SVM obtuvo los mejores resultados.

5. Conclusiones

En este estudio, exploramos tres diferentes técnicas de sobremuestreo con la finalidad de balancear el *dataset* multiclase desbalanceado SGB y obtener mediante tres clasificadores, modelos predictivos utilizando datos balanceados. Buscamos comprobar si balancear los datos mejoran el rendimiento de los modelos.

Se obtuvieron cuatro *subsets* a través del método de binarización "uno contra todos". A cada *subset* de entrenamiento aplicamos los métodos MWMOTE, RWO, y RACOG para sobremuestrear la clase minoritaria y equilibrarla con la clase mayoritaria. Por otro lado, se crearon modelos predictivos utilizando datos desbalanceados. Utilizamos 3 clasificadores que determinan sus resultados de diferente manera: C4.5 (jerárquico), SVM lineal (kernel) y JRip (reglas). La medida de rendimiento utilizada fue la Curva ROC y el área bajo la curva AUC. Aplicamos la prueba Wilcoxon para conocer si existe diferencia estadísticamente significativa entre los modelos balanceados y los modelos desbalanceados, siempre y cuando los datos balanceados sean superiores a los datos desbalanceados.

El método de sobremuestreo RWO obtuvo los mejores resultados. Aplicando RWO obtuvimos 12 modelos, en ocho ocasiones superaron los datos desbalanceados, de los cuales, 4 de ellos obtuvieron diferencia estadísticamente significativa; en 4 modelos tuvieron bajo rendimiento respecto a los datos desbalanceados. Usando MWMOTE, 7 modelos superaron los datos desbalanceados, en 3 ocasiones se encontró diferencia estadísticamente significativa; en 5 modelos obtuvieron mejor rendimiento los datos desbalanceados. El método RACOG obtuvo el peor rendimiento. En 7 modelos que superaron los datos desbalanceados, solo uno de ellos obtuvo una diferencia estadísticamente significativa; en 5 modelos obtuvieron mejor rendimiento los datos desbalanceados.

Respecto a los clasificadores, SVM obtuvo el mejor rendimiento. Utilizando SVM se crearon 12 modelos balanceados, 10 modelos superaron a los modelos desbalanceados, en 4 de ellos se encontró diferencia estadísticamente significativa; 2 modelos fueron superados por los datos desbalanceados. Utilizando el clasificador JRip, ocho modelos superaron los modelos desbalanceados, sin embargo, solo dos de ellos se encontró diferencia estadísticamente significativa; cuatro modelos fueron superados por los modelos desbalanceados. El clasificador C4.5 obtuvo el peor rendimiento. Aplicando C4.5 se obtuvieron 12 modelos balanceados, cuatro de ellos superaron los modelos desbalanceados, sin embargo, solo dos tuvieron diferencia estadísticamente significativa; ocho modelos balanceados fueron superados por los modelos desbalanceados.

Los resultados revelan que balancear el conjunto de datos optimiza el rendimiento de los modelos predictivos. Sobremuestrear la clase minoritaria utilizando el método RWO en combinación con el clasificador SVM obtuvo los mejores resultados.

Este estudio es una serie de trabajos que nos permitirán encontrar el mejor método de balanceo a nivel de datos, utilizando sobremuestreo y submuestreo, que incremente el rendimiento de los modelos predictivos para identificar el subtipo de SGB. Esto podrá ser una herramienta complementaria para los especialistas y les permita ayudar en la identificación de alguna de las variantes de SGB que padezca el paciente.

Como trabajos futuros, experimentaremos con variantes clásicas de SMOTE, hibridación de diferentes técnicas de muestreo. Aplicaremos diferentes clasificadores individuales y clasificadores combinados para mejorar el rendimiento de los modelos actuales.

Referencias

1. Barua, S., Islam, M.M., Yao, X., Murase, K.: MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning. IEEE Transactions on Knowledge and Data Engineering 26(2), 405–425 (2014)
2. Canul-Reich, J., Frausto-Solís, J., Hernández-Torruco, J.: A predictive model for guillain-barré syndrome based on single learning algorithms. Computational and Mathematical Methods in Medicine pp. 1–9 (2017)

3. Canul-Reich, J., Hernández-Torruco, J., Chávez-Bosquez, O., Hernández-Ocaña, B.: A predictive model for guillain–barré syndrome based on ensemble methods. Computational Intelligence and Neuroscience pp. 1–10 (2018)
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
5. Cordón, I., García, S., Fernández, A., Herrera, F.: imbalance: Preprocessing Algorithms for Imbalanced Datasets (2018)
6. Das, B., Krishnan, N.C., Cook, D.J.: RACOG and wRACOG: Two probabilistic oversampling techniques. IEEE Transactions on Knowledge and Data Engineering 27(1), 222–234 (2015)
7. Esposito, S., Longo, M.R.: Guillain–barré syndrome. Autoimmunity Reviews 16(1), 96–101 (2017)
8. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F.: Learning from Imbalanced Data Sets. Springer International Publishing (2018)
9. Hernández-Torruco, J., Canul-Reich, J., Frausto-Solís, J., Méndez-Castillo, J.J.: Feature selection for better identification of subtypes of guillain-barré syndrome. Computational and Mathematical Methods in Medicine 2014, 1–9 (2014)
10. Karalok, Z.S., Taskin, B.D., Yanginlar, Z.B., Gurkas, E., Guven, A., Degerliyurt, A., Unlu, E., Kose, G.: Guillain-barré syndrome in children: subtypes and outcome. Child's Nervous System 34(11), 2291–2297 (2018)
11. Loyola-González, O., Medina-Pérez, M.A., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., Monroy, R., García-Borroto, M.: PBC4cip: A new contrast pattern-based classifier for class imbalance problems. Knowledge-Based Systems 115, 100–109 (2017)
12. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F.: e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien (2018)
13. Ng, S., Tse, P., Tsui, K.: A one-versus-all class binarization strategy for bearing diagnostics of concurrent defects. Sensors 14(1), 1295–1321 (2014)
14. Shatnawi, R., Li, W., Swain, J., Newman, T.: Finding software metrics threshold values using ROC curves. Journal of Software Maintenance and Evolution: Research and Practice 22(1), 1–16 (2009)
15. Soto-Hernández, J.L., de León Rosales, S.P., Cañas, E.S.V., Cárdenas, G., Loza, K.C., Díaz-Quiñonez, J.A., López-Martínez, I., Jiménez-Corona, M.E., Ruiz-Matus, C., Morales, P.K.: Guillain-Barré syndrome associated with Zika virus infection: A prospective case series from Mexico. Frontiers in Neurology 10 (2019)
16. Thakur, S., Meenakshi, E., Priya, A.: Detection of malicious URLs in big data using RIPPER algorithm. In: IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology RTEICT (2017)
17. Torres-Vásquez, M., Chávez-Bosquez, O., Hernández-Ocaña, B., Hernández-Torruco, J.: Balanceo de datos del sındrome de guillain-barré utilizando smote para la clasificación de subtipos. Research in Computing Science 148(7), 113–125 (2019)
18. Torres-Vásquez, M., Chávez-Bosquez, O., Hernández-Ocaña, B., Hernández-Torruco, J.: Classification of guillain–barré syndrome subtypes using sampling techniques with binary approach. Symmetry 12(3), 482 (2020)
19. Wijdicks, E.F., Klein, C.J.: Guillain-barré syndrome. Mayo Clinic Proceedings 92(3), 467–479 (2017)
20. Witten, I.H., Frank, E., Hall, M.A., Pañ, C.: Data Mining, Practical Machine Learning Tools and Techniques. Morgan Kaufmann (2017)

21. Zhang, H., Li, M.: RWO-sampling: A random walk over-sampling approach to imbalanced data classification. Information Fusion 20, 99–116 (2014)