

Balanced data of the Guillain-Barré Syndrome using SMOTE for the classification of subtypes

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

Universidad Juárez Autónoma de Tabasco, División Académica de Informática y
Sistemas, Cunduacán, Tabasco, México

181H18001@alumno.ujat.mx, {jose.hernandezt,betania.hernandez,
oscar.chavez}@ujat.mx

Resumen. El trastorno neurológico llamado Síndrome de Guillain-Barré (SGB) afecta principalmente al sistema nervioso periférico y se caracteriza por un rápido desarrollo, en algunos casos es mortal. Los subtipos del SGB son principalmente Polineuropatía Desmielinizante Inflamatoria Aguda (AIDP), Neuropatía Axonal Motora Aguda (AMAN), Neuropatía Axonal Sensorial Aguda (AMSAN) y Síndrome de Miller-Fisher (MF). Cada subtipo debe tratarse de manera diferente, el tratamiento y los costos varían según el subtipo desarrollado. Por lo tanto, es esencial diagnosticar temprano el subtipo de SGB que sufre el paciente. El conjunto de datos incluye 129 registros de pacientes mexicanos. Este conjunto de datos se ha utilizado para construir una serie de modelos predictivos utilizando clasificadores simples y combinados Debido a que los datos originales del SGB están desbalanceados, balanceamos los datos a través del sobremuestreo sintético y utilizamos clasificadores simples para mejorar los modelos predictivos previos. En este estudio, usamos la Técnica de Sobremuestreo de Minorías Sintéticas (SMOTE) para balancear los datos, y aplicamos tres clasificadores simples con diferentes enfoques: C4.5, SVM, JRip. Utilizamos la técnica OvO (*One for One*) para obtener 6 subproblemas binarios de subtipos SGB. Usamos la curva ROC (*Receiver Operating Characteristic*) y la prueba de Wilcoxon para evaluar el rendimiento de los modelos obtenidos con datos desbalanceados en comparación con los modelos obtenidos con los datos balanceados para cada caso. Los resultados mostraron que la técnica SMOTE combinada con el clasificador JRip obtuvo los mejores resultados. Sin embargo, en algunos casos, los resultados con datos desbalanceados mejoraron los resultados utilizando datos balanceados.

Palabras clave: desbalanceo de datos, clasificación multiclase, SVM, JRip, C4.5, curva ROC, prueba Wilcoxon.

Balancing Guillain-Barré Data Using SMOTE for Subtype Classification

Abstract. The neurological disorder called Guillain-Barré Syndrome (GBS) mainly affects the peripheral nervous system and it is charac-

terized by a fast development, in some cases is mortal. The GBS subtypes are mainly Acute Inflammatory Demyelinating Polyneuropathy (AIDP), Acute Motor Axonal Neuropathy (AMAN), Acute Motor Sensory Axonal Neuropathy (AMSAN), and Miller-Fisher Syndrome (MF). Each subtype must be treated differently, treatment and costs differ according to the subtype developed by the patient. So, it is essential to early diagnose the subtype of GBS that the patient suffers. The dataset includes 129 records of Mexican patients. This dataset has been used to build a number of predictive models using single classifiers and ensemble methods. As the original GBS data is imbalanced, we balanced the data through synthetic oversampling and we used single classifiers to possible improve previous predictive models. In this study, we use SMOTE (Synthetic Minority Over-sampling Technique) for balancing data and we applied three single classifiers with different approaches: C4.5, SVM, JRip. We used OvO (One versus One) technique to obtain 6 binary subproblems of SGB subtypes. The ROC (Receiver Operating Characteristic) curve and Wilcoxon test both evaluated performance of the imbalanced models versus better balanced data for each case. Results showed that the SMOTE technique combined with the JRip classifier obtained the best results. However, some cases, results with imbalanced data improved results using balanced data.

Keywords: imbalance data, multiclass classification, SVM, JRip, C4.5, ROC curve, Wilcoxon test.

1. Introducción

El Síndrome de Guillain-Barré (SGB) es un trastorno autoinmune en el que el sistema inmunológico del cuerpo ataca los nervios, causando debilidad muscular y, a veces, parálisis. Mientras que la mayoría de las personas eventualmente se recuperan completamente del SGB, otras se quedan con daño nervioso permanente. En raras ocasiones el SGB puede llevar a la muerte. Algunas de sus principales características son [21]:

- La causa es desconocida, pero está asociada con infección respiratoria o gastrointestinal.
- Causa serios daños a la habilidad motora.
- Afecta de 1.3 a 2 por cada 100,000 personas anualmente.
- Tiene una tasa de mortalidad del cinco al quince por ciento.

Los cuatro subtipos principales del SGB son: Polineuropatía Desmielinizante Inflamatoria Aguda (AIDP, *Acute Inflammatory Demyelinating Polyneuropathy*), Neuropatía Axonal Motora Aguda (AMAN, *Acute Motor Axonal Neuropathy*), Neuropatía Axonal Sensorial Aguda (AMSAN, *Acute Motor Sensory Axonal Neuropathy*) y Síndrome de Miller-Fisher (MF). En la Tabla 1 se describen las características de cada uno de los subtipos de SGB.

Es esencial diagnosticar el subtipo de SGB que el paciente sufre de manera temprana. Cada subtipo debe tratarse de manera diferente, el tratamiento y

Tabla 1. Características de los subtipos del SGB.

Polineuropatía Desmielinizante Inflamatoria Aguda		
Características patológicas	Características clínicas	Enfermedades previas
Desmielinización periférica multifocal. Remielinización lenta. Probablemente mecanismos inmunitarios tanto humorales como celulares.	El subtipo más común (hasta el 90 por ciento de los casos de SGB en los Estados Unidos). Debilidad progresiva, simétrica; hiporreflexia o arreflexia.	Virus del herpes. Campylobacter jejuni. Micoplasma. Neumonía.
Neuropatía Axonal Motora Aguda		
Características patológicas	Características clínicas	Enfermedades previas
Anticuerpos contra los gangliósidos GM1, GD1a, GalNAc-GD1a y GD1b en los axones del nervio motor periférico; no desmielinización.	La neuropatía axonal aguda del motor o del sensor sensorial representa del 5 al 10 por ciento de los casos de SGB. Fuertemente asociado con Campylobacter jejuniinfection; Más común en el verano, en pacientes más jóvenes y en China o Japón. Sólo síntomas motores. Los reflejos profundos del tendón pueden ser preservados.	Campylobacter jejuni.
Neuropatía Axonal Sensorial Aguda		
Características patológicas	Características clínicas	Enfermedades previas
Mecanismo similar al de la neuropatía axonal motora aguda, pero con degeneración axonal sensorial.	Similar a los de la neuropatía axonal motora aguda, pero con afectación predominantemente sensorial.	Campylobacter jejuni.
Síndrome Miller Fisher		
Características patológicas	Características clínicas	Enfermedades previas
Desmielinización. Anticuerpos de inmunoglobulina G contra los gangliósidos GQ1b, GD3 y GT1a.	Raro (3 por ciento de los casos de SGB en los Estados Unidos) Oftalmoplejia bilateral. Ataxia. Areflexia. Facial, debilidad bulbar ocurre en el 50 por ciento de los casos. Tronco, la debilidad de la extremidad se produce en el 50 por ciento de los casos.	Ocasionalmente, Campylobacter jejuni, la mayoría de los casos con historia desconocida o Mycoplasma pneumoniae.

los costos varían según el subtipo desarrollado por el paciente. El método de diagnóstico tradicional de los subtipos SGB es a través de estudios fisiológicos y de laboratorio, que son tardados y costosos.

Hoy en día, el aprendizaje automático se utiliza ampliamente para diagnosticar enfermedades, como el cáncer y el Alzheimer. Sin embargo, un número limitado de trabajos aborda el SGB. En [16] se aplicó la regresión logística multinomial para identificar los predictores clínicos y electrodiagnósticos de la parálisis respiratoria neuromuscular en el SGB. Los predictores encontrados fueron la progresión temprana hacia discapacidad pico, la disfunción bulbar y la inestabilidad autonómica. En otro trabajo [13], aplicaron regresión logística para identificar las variables clínicas asociadas con la necesidad de ventilación mecánica en los pacientes con SGB. Las variables identificadas fueron el inicio simultáneo de la debilidad motora de las extremidades superiores e inferiores como síntoma inicial, el grado de potencia de la extremidad superior $<3/5$ en el nadir y la debilidad bulbar. En [19] se aplicó análisis discriminante lineal disperso para optimizar la precisión electrodiagnóstica de los subtipos del SGB.

En este estudio, nuestro objetivo fue investigar si a través de técnicas de balanceo de datos es posible mejorar los modelos predictivos para el SGB creados anteriormente, donde el conjunto de datos estaba desbalanceado. Estos modelos se crearon utilizando clasificadores simples y combinados, utilizando los datos desbalanceados originales [1,2]. Dado que este es un estudio exploratorio, para balancear los datos utilizamos SMOTE (*Synthetic Minority Over-sampling TEchnique*) [3], una técnica ampliamente utilizada con resultados probados.

Aplicamos tres clasificadores simples con diferentes enfoques: C4.5 (árbol de decisión, basada en instancias), SVM (*Support Vector Machine*) basado en kernel, y JRip (basado en reglas). Comparamos su rendimiento en experimentos que consisten en balancear los datos al 100 %, 200 %, 300 %, 400 %, 500 %, 1000 % para 6 subproblemas binarios de subtipos SGB utilizando el enfoque OvO (*One versus One*). El rendimiento de los modelos se evaluó mediante la curva de características operativas del receptor (ROC del inglés *Receiver Operating Characteristic*), medida típica para conjuntos de datos desbalanceados. Utilizamos la prueba no paramétrica de Wilcoxon [5] para investigar una diferencia estadística entre los modelos desbalanceados y balanceados.

Este artículo está organizado de la siguiente forma: En la sección 2 presentamos una descripción del conjunto de datos, los algoritmos de aprendizaje automático y la medida de rendimiento utilizada en el estudio. La sección 3 describe el procedimiento experimental. En la sección 4 mostramos y discutimos los resultados experimentales. Finalmente, en la sección 5 resumimos los resultados, damos conclusiones del estudio y también sugerimos algunos trabajos futuros.

2. Materiales y métodos

2.1. Conjunto de datos

Los datos se recopilaron en el Instituto Nacional de Neurología y Neurocirugía (México) de 1993 a 2002. Se tienen registros de 129 pacientes, cada uno

clasificado con uno de los subtipos del SGB: 20 AIDP, 37 AMAN, 59 AMSAN, 13 Miller-Fisher.

De acuerdo con un estudio anterior [9], seleccionamos los atributos relevantes que se muestran en la Tabla 2. Las primeras cuatro variables son valores enteros, las siguientes son reales.

Tabla 2. Variables utilizadas en este trabajo.

Etiqueta	Nombre de las variables
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

SMOTE. Utilizamos la técnica de preprocesamiento SMOTE para sobremuestrear la clase minoritaria hasta igualar la clase mayoritaria de nuestro conjunto de datos desbalanceado.

En [3] se introdujo uno de los métodos de muestreo más utilizados y con mayor éxito en casos de problemas de desbalanceo de clases binario, la Técnica de Sobremuestreo de Minorías Sintéticas: SMOTE (*Synthetic Minority Over-sampling TEchnique*). Esta técnica sobremuestra la clase minoritaria al crear datos sintéticos o artificiales basado en las similitudes del espacio de características entre los ejemplos de minoría existentes, a diferencia de otras técnicas que solo reemplazan los datos en forma aleatoria. SMOTE toma la clase minoritaria y la sobremuestra tomando cada muestra de clase minoritaria e introduciendo ejemplos sintéticos a lo largo de los segmentos de línea que unen a cualquiera de los vecinos más cercanos a la clase k minoritaria. Dependiendo de la cantidad de sobremuestreo requerido, los vecinos de los vecinos más cercanos se eligen al azar. Estos nuevos datos creados en forma sintética mejoran las técnicas

anteriores que reemplazan el sobremuestreo en forma simple. Los datos sintéticos aumentan el conjunto de datos original ayudando al clasificador a crear regiones de decisión mas grande y menos específicas logrando una mejora significativa en el clasificador [6].

El algoritmo SMOTE realiza los siguientes pasos:

- Recibe como parámetro el porcentaje de ejemplos a sobremuestrear.
- Calcula el número de ejemplos que tiene que generar.
- Calcula los k vecinos más cercanos de los ejemplos de la clase minoritaria.
- Genera los ejemplos siguiendo este proceso:
 - Para cada ejemplo de la clase minoritaria, elige aleatoriamente el vecino a utilizar para crear el nuevo ejemplo.
 - Para cada atributo del ejemplo a sobremuestrear, calcula la diferencia entre el vector de atributos muestra y el vecino elegido.
 - Multiplica esta diferencia por un número aleatorio entre 0 y 1.
 - Suma este último valor al valor original de la muestra.
 - Devuelve el conjunto de ejemplos sintéticos.

En la Figura 1 podemos observar el funcionamiento de SMOTE. Este crea objetos sintéticos en la clase minoritaria mediante la interpolación de un objeto y sus k vecinos más cercanos. En la Figura 1a podemos observar el conjunto de datos formado por dos clases una clase mayoritaria y otra minoritaria. En la Figura 1b muestra los vecinos más cercanos seleccionados para aplicar SMOTE. También se observan las instancias creadas en forma sintética de la clase minoritaria. En la Figura 1c se muestra el conjunto de datos balanceados a través del sobremuestreo.

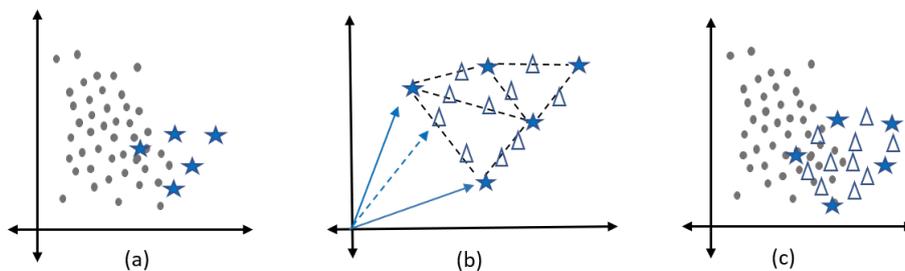


Fig. 1. Generación de datos sintéticos con SMOTE.

Clasificadores simples C4.5 (árbol de decisión, basada en instancias), Máquina de Soporte Vectorial (SVM del inglés *Support Vector Machine*, basado en kernel), JRip (basado en reglas). Utilizamos estos clasificadores para discriminar las instancias de cada clase y comparar los resultados.

2.3. Medidas de rendimiento

Utilizamos la curva ROC, como medida de rendimiento para el conjunto de datos desbalanceado. Es un gráfico bidimensional que proporciona un buen resumen del rendimiento de un modelo de clasificación en presencia de conjuntos de datos desbalanceados con costos de error desiguales [8].

El área de la gráfica tiene un valor entre 0.5 y 1, donde un valor de 1 representa un diagnóstico perfecto y 0.5 representa una prueba sin diagnóstico de capacidad discriminatoria.

2.4. Técnica de binarización

Dividimos el conjunto de datos en subproblemas binarios utilizando el enfoque OvO. Esta técnica entrena un clasificador para cada par de clases posibles $(n-1)/2$ (aprendizaje por pares), por lo tanto, obtuvimos seis pares de conjuntos de datos desbalanceados con el mismo número de combinaciones.

Utilizamos la técnica de binarización OvO, ampliamente utilizada en problemas de clasificación. OvO tiene un mejor desempeño al aplicar la prueba de Wilcoxon en comparación con la técnica OvA [7]. Desde una perspectiva médica, OvO puede ayudar a los médicos a distinguir un subtipo de otro.

2.5. Validación

Utilizamos la evaluación *train-test* para cada clasificador individual, empleando dos tercios de los datos para el entrenamiento y un tercio para las pruebas.

3. Procedimiento experimental

Comenzamos creando 6 conjuntos de datos binarios a partir del conjunto de datos original utilizando OvO:

- SGB1 (57 instancias): AIDP (20 instancias) vs AMAN (37 instancias).
- SGB2 (79 instancias): AIDP (20 instancias) vs AMSAN (59 instancias).
- SGB3 (33 instancias): AIDP (20 instancias) vs MF (13 instancias).
- SGB4 (96 instancias): AMAN (37 instancias) vs AMSAN (59 instancias).
- SGB5 (50 instancias): AMAN (37 instancias) vs MF (13 instancias).
- SGB6 (72 instancias): AMSAN (59 instancias) vs MF (13 instancias).

Luego, dividimos cada uno de los conjuntos de datos en dos conjuntos, 66 % para los datos de entrenamiento y 34 % para los datos de prueba. Balanceamos los datos de entrenamiento utilizando SMOTE. La clase minoritaria de cada conjunto de entrenamiento se sobremuestrea al 100 %, 200 %, 300 %, 400 %, 500 %, 1000 %, según la literatura [3].

La Tabla 3 muestra el número de instancias de la clase minoritaria para cada conjunto de datos binarios. La primera columna muestra los seis casos de SGB n formados por la técnica OvO. La segunda columna muestra las instancias

originales correspondientes a la clase minoritaria del conjunto de datos original para cada caso. En la tercera columna, se puede ver el número de instancias de entrenamiento para cada caso. En las columnas 4 a 9, el resultado del sobremuestreo de las instancias de entrenamiento se muestra en 100 %, 200 %, 300 %, 400 %, 500 % y 1000 % consecutivamente.

Tabla 3. Sobremuestreo de la clase minoritaria utilizando SMOTE.

SGB	Original	Entrenamiento	100 %	200 %	300 %	400 %	500 %	1000 %
SGB1	20	11	22	33	44	55	66	121
SGB2	20	14	28	42	56	70	84	154
SGB3	13	9	18	27	36	45	54	99
SGB4	37	24	48	72	96	120	144	264
SGB5	13	9	18	27	36	45	54	99
SGB6	13	10	20	30	40	50	60	110

Para cada conjunto de datos SGB_n , calculamos el promedio del resultado de la curva ROC de 60 ejecuciones independientes con diferentes semillas utilizando los datos desbalanceados y balanceados al 100 %, 200 %, 300 %, 400 %, 500 % y 1000 % para los tres clasificadores. Comparamos en cada caso el promedio del modelo desbalanceado contra el mejor promedio del modelo balanceado. Realizamos la prueba no paramétrica de Wilcoxon [5] para investigar una diferencia estadística entre los modelos desbalanceados y balanceados. Utilizamos un valor de significación de 0.05. Utilizamos pruebas no paramétricas porque las condiciones iniciales que garantizan la credibilidad de las pruebas paramétricas pueden no cumplirse, lo que hace que el análisis estadístico sea menos confiable con este tipo de pruebas [7].

Los experimentos de R se realizaron en RStudio 1.1.423, con SMOTE del paquete DMwR 0.4.1 [18]. Utilizamos el paquete RWeka 0.4-39 [22], incluidos los clasificadores C4.5 y JRip. Clasificadores SVM del paquete e1071 1.7-0 [11]. Otros paquetes utilizados fueron rJava [20], caret [10], rpart [17], imbalance [4], Proc [14], lattice [15], y rpart.plot [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 4 muestra la comparación de los resultados de la clasificación utilizando un conjunto de datos desbalanceados con los resultados de la clasificación utilizando el conjunto de datos balanceados al 100 %, 200 %, 300 %, 400 %, 500 % y 1000 % para cada clasificador, ROC fue la métrica utilizada. Cada valor es el promedio de los resultados obtenidos en las 60 ejecuciones. El mejor resultado de cada clasificador se resalta en negrita. La primera columna muestra los casos de SGB y su número de instancias. La última columna muestra el resultado de la

Tabla 4. Resultados de clasificación promedio de datos desbalanceados contra datos balanceados.

Subtipo (instancias)	Clasificador	Datos des-		Balanceo (%)						Resultados de prueba Wilcoxon
		balanceados		100	200	300	400	500	1000	
		ROC	ROC	ROC	ROC	ROC	ROC	ROC	ROC	
SGB1	C4.5	0.9563	0.9479	0.9438	0.9493	0.9528	0.9556	0.9576	NS	
AIDP-AMAN (20-37)	SVM	0.9618	0.9618	0.9632	0.9639	0.9625	0.9632	0.9632	NC	
	JRip	0.9507	0.9458	0.9424	0.9403	0.9382	0.9319	0.9389	NA	
SGB2	C4.5	0.8656	0.8430	0.8485	0.8375	0.8502	0.8607	0.8551	NA	
AIDP-AMSAN (20-59)	SVM	0.9618	0.9618	0.9632	0.9639	0.9625	0.9632	0.9632	NS	
	JRip	0.8472	0.8455	0.8285	0.8480	0.8561	0.8480	0.8549	NS	
SGB3	C4.5	0.8132	0.8313	0.7986	0.7889	0.7729	0.7958	0.7965	NS	
AIDP-MF (20-13)	SVM	0.7097	0.6306	0.6486	0.6472	0.6465	0.6563	0.6535	NA	
	JRip	0.8458	0.8250	0.7778	0.7750	0.7646	0.7292	0.7382	NA	
SGB4	C4.5	0.9132	0.9079	0.9096	0.9172	0.9062	0.9207	0.9093	NS	
AMAN-AMSAN (37-59)	SVM	0.8863	0.8836	0.8844	0.8843	0.8840	0.8821	0.8827	NA	
	JRip	0.8809	0.8955	0.9042	0.9019	0.9043	0.9071	0.9065	S	
SGB5	C4.5	0.8736	0.8868	0.8792	0.8833	0.8701	0.8861	0.8868	NS	
AMAN-MF (37-13)	SVM	0.8910	0.8611	0.8715	0.8792	0.8840	0.8847	0.8847	NA	
	JRip	0.8799	0.9014	0.8799	0.8875	0.8903	0.8861	0.8889	NS	
SGB6	C4.5	0.8007	0.8697	0.8185	0.8287	0.8084	0.8041	0.7839	S	
AMSAN-MF (59-13)	SVM	0.7388	0.7882	0.7646	0.7651	0.7522	0.7469	0.7534	S	
	JRip	0.8430	0.8720	0.8393	0.8306	0.8213	0.8111	0.8061	S	

prueba de Wilcoxon, donde *NS* no representa una diferencia significativa entre los resultados que utilizan datos desbalanceados y el mejor resultado después de balancear los datos; *NC* representa que la prueba no se pudo realizar debido a la similitud entre los datos comparados; *NA* significa que no se realizó ninguna prueba porque el mejor valor se obtuvo utilizando datos desbalanceados; **S** representa que los resultados son estadísticamente significativos.

Para SGB1, dos clasificadores mejoraron el rendimiento utilizando datos balanceados: C4.5 con un valor ROC de 0.9576 y balanceo de 1000 %, y SVM con valor ROC de 0.9639 y un balanceo de 300 %. En el primer caso, no hubo diferencia significativa con respecto a los datos desbalanceados, según la prueba de Wilcoxon. En el segundo caso, no se pudo realizar la prueba de Wilcoxon debido que los resultados obtenidos con el conjunto de datos desbalanceado y el balanceado al 300 % fueron similares. En el tercer caso, JRip no obtuvo un mejor rendimiento con los datos balanceados en ningún caso.

Para SGB2, dos clasificadores mejoraron el rendimiento utilizando datos balanceados: SVM con un valor ROC de 0.9639 y un balanceo del 300 %, y JRip con un valor ROC de 0.8561 y un balanceo de 400 %. En el primer caso, el clasificador C4.5 obtuvo un mejor rendimiento con datos desbalanceados con respecto a los datos balanceados. En el segundo y tercer caso, la prueba de Wilcoxon no encontró diferencia significativa con respecto a los datos desbalanceados.

En SGB3, C4.5 mejoró el rendimiento utilizando datos balanceados con un valor de ROC 0.8313 y un balanceo del 100 %. En el primer caso, la prueba de

Wilcoxon no encontró diferencia significativa con respecto a los datos desbalanceados. En el segundo y tercer caso, no hubo mejor rendimiento para SVM y JRip utilizando datos balanceados.

Para SGB4, dos clasificadores mejoraron el rendimiento utilizando datos balanceados: C4.5 con un valor ROC de 0.9207 y un balanceo del 500 %, y JRip con un valor ROC de 0.9071 y un balanceo del 500 %. En el primer caso, la prueba de Wilcoxon no encontró diferencia significativa con respecto a los datos desbalanceados. En el segundo caso, SVM no obtuvo un mejor rendimiento utilizando datos balanceados. En el tercer caso, la prueba de Wilcoxon encontró una diferencia significativa entre los datos desbalanceados y balanceados.

En SGB5, dos clasificadores mejoraron el rendimiento utilizando datos balanceados: C4.5 con un valor ROC de 0.8868 y un balanceo del 100 %, y JRip con un valor ROC de 0.9014 y un balanceo del 100 %. En el primer y tercer caso, la prueba de Wilcoxon no encontró diferencia significativa con respecto a los datos desbalanceados. En el segundo caso, no hubo mejor rendimiento utilizando datos balanceados.

Para SGB6, los tres clasificadores mejoraron el rendimiento utilizando datos balanceados: C4.5 con un valor ROC de 0.8697 y un balanceo del 100 %, SVM con un valor ROC de 0.7882 y un balanceo del 100 % y JRip con un valor ROC de 0.8720 y un balanceo del 100 %. La prueba de Wilcoxon encontró una diferencia significativa entre los datos desbalanceados y los datos balanceados en los tres casos.

Los tres clasificadores mejoraron el rendimiento en 6 casos. C4.5 mejoró un caso para datos desbalanceados y cinco casos utilizando datos balanceados. Sin embargo, solo un caso obtuvo diferencia significativa. SVM mejoró tres casos para datos desbalanceados y tres utilizando datos balanceados. Sin embargo, solo un caso obtuvo diferencia significativa. JRip mejoró dos casos para datos desbalanceados y cuatro casos con datos balanceados, dos casos con diferencias significativas. Todos los casos mejorados de los tres clasificadores se obtuvieron con datos balanceados.

Se obtuvo un mejor rendimiento en 12 casos utilizando datos balanceados y 6 casos obtuvieron un mejor rendimiento con el conjunto de datos desbalanceados. El balanceo al 100 % obtuvo el mayor rendimiento (seis casos, tres de ellos con una diferencia significativa). El balanceo a 200 % no obtuvo un mejor rendimiento. El balanceo a 300 % mejoró el rendimiento en dos casos (sin una diferencia significativa). El balanceo a 400 % mejoró el rendimiento en un caso (sin una diferencia significativa). El balanceo en 500 % mejoró el rendimiento en dos casos (una diferencia significativa). El balanceo a 1000 % mejoró el rendimiento en un caso (sin una diferencia significativa).

Estos resultados podrían ser un indicio de que el número de instancias influye en la clasificación una vez que los datos están balanceados. Se requieren más experimentos para confirmarlo.

5. Conclusiones

En este trabajo nos propusimos investigar si es posible mejorar los modelos predictivos para los subtipos de SGB creados en un estudio previo, mediante el uso de técnicas de balanceo de datos. Para balancear los datos utilizamos SMOTE, ya que es una técnica ampliamente utilizada con resultados probados. Este trabajo constituye el primer esfuerzo en el uso de técnicas de balanceo de datos para mejorar los modelos predictivos de SGB.

Para este primer estudio, seleccionamos tres clasificadores simples con diferentes enfoques: C4.5 (árbol de decisión, basada en instancias), Máquina de Soporte Vectorial (SVM del inglés *Support Vector Machine*, basado en kernel), JRip (basado en reglas). Comparamos su rendimiento en experimentos que consistían en balancear los datos al 100 %, 200 %, 300 %, 400 %, 500 %, 1000 % para 6 subproblemas binarios de subtipos de SGB (OvO).

El mejor resultado se obtuvo en el conjunto de datos SGB6 ya que los tres clasificadores crearon modelos predictivos utilizando datos balanceados que mejoraron significativamente los modelos con datos desbalanceados. Este resultado se logró al balancear los datos al 100 %.

En cuanto al rendimiento de los clasificadores, JRip fue el mejor, ya que se obtuvo modelos predictivos utilizando datos balanceados que mejoraron significativamente los modelos con datos desbalanceados. Esto mejoró los modelos en dos casos (SGB4 y SGB6). En el nivel de balanceo de datos, llegamos a la conclusión de que los mejores resultados se obtuvieron con un balance del 100 %, ya que en seis casos los modelos con datos balanceados superaron a los modelos con datos desbalanceados, tres de ellos con una diferencia significativa. Observamos que los casos con el mayor número de instancias obtuvieron el mejor rendimiento: SGB4 con 37 instancias para la clase minoritaria y 59 para la clase mayoritaria, y SGB6 con 13 instancias para la clase minoritaria y 59 para la clase mayoritaria.

Los resultados logrados en este estudio muestran que es factible construir un modelo predictivo para apoyar a los especialistas en la predicción automática que permita la distinción entre las diferentes variantes de esta enfermedad. La gravedad y el tratamiento del SGB varía entre los diferentes subtipos, por lo que la diferenciación entre ellos ayuda a la correcta identificación del subtipo que padece el paciente y de esta forma los especialistas pueden tomar las medidas necesarias para atenderlo adecuadamente.

Desde el punto de vista del aprendizaje automático, encontramos que SMOTE nos permitió crear algunos modelos con datos balanceados que superaron a los modelos con datos desbalanceados. Lo anterior representa una contribución en la exploración del rendimiento de las técnicas de balanceo de datos usando conjuntos de datos reales. Como trabajo futuro, experimentaremos con más variantes de la técnica SMOTE, y también con diferentes clasificadores simples y combinados para construir modelos predictivos más precisos.

Agradecimientos. Esta publicación fue financiada con recursos del Programa de Fortalecimiento de la Calidad Educativa (PFCE) 2019, bajo el proyecto

número P/PFCE-2019-27MSU0018V-11. Queremos agradecer al CONACYT por el apoyo al Doctorado en Ciencias de la Computación de la Universidad Juárez Autónoma de Tabasco.

Referencias

1. 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* 2017, 1–9 (2017)
2. 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* 2018, 1–10 (2018)
3. 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)
4. Córdón, I., García, S., Fernández, A., Herrera, F.: *imbalanced: Preprocessing Algorithms for Imbalanced Datasets* (2018), <https://CRAN.R-project.org/package=imbalanced>, r package version 1.0.0
5. Cuzick, J.: A wilcoxon-type test for trend. *Statistics in Medicine* 4(4), 543–547 (1985)
6. Fernandez, A., Garcia, S., Herrera, F., Chawla, N.V.: SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research* 61, 863–905 (2018)
7. Fernández, A., López, V., Galar, M., del Jesus, M.J., Herrera, F.: Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-Based Systems* 42, 97–110 (2013)
8. Gu, Q., Zhu, L., Cai, Z.: Evaluation measures of the classification performance of imbalanced data sets. In: *Communications in Computer and Information Science*, pp. 461–471. Springer Berlin Heidelberg (2009)
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. from Jed Wing, M.K.C., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., Hunt, T.: *caret: Classification and Regression Training* (2018), <https://CRAN.R-project.org/package=caret>, R package version 6.0-81
11. 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), <https://CRAN.R-project.org/package=e1071>, R package version 1.7-0
12. Milborrow, S.: *rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'* (2018), <https://CRAN.R-project.org/package=rpart.plot>, r package version 3.0.5
13. Paul, B., Prasad, K., Singh, M., Tripathi, M., Bhatia, R., Padma, M.: Clinical predictors of mechanical ventilation in Guillain-Barré syndrome. *Neurology India* 60(2), 150 (2012)
14. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., Muller, M.: *proc: an open-source package for r and s+ to analyze and compare roc curves*. *BMC Bioinformatics* 12, 77 (2011)

15. Sarkar, D.: *Lattice: Multivariate Data Visualization with R*. Springer, New York (2008), <http://lmdvr.r-forge.r-project.org>, ISBN 978-0-387-75968-5
16. Sundar, U., Abraham, E., Gharat, A., Yeolekar, M., Trivedi, T., Dwivedi, N.: Neuromuscular respiratory failure in Guillain-Barré syndrome: Evaluation of clinical and electrodiagnostic predictors. *Journal of Association of Physicians of India* 53(1), 764–768 (2005)
17. Therneau, T., Atkinson, B.: *rpart: Recursive Partitioning and Regression Trees* (2018), <https://CRAN.R-project.org/package=rpart>, r package version 4.1-13
18. Torgo, L.: *Datamining with R learning with case studies*. Chapman & Hall/CRC (2011)
19. Uncini, A., Ippoliti, L., Shahrizaila, N., Sekiguchi, Y., Kuwabara, S.: Optimizing the electrodiagnostic accuracy in Guillain-Barré syndrome subtypes: Criteria sets and sparse linear discriminant analysis. *Clinical Neurophysiology* 128(7), 1176–1183 (2017)
20. Urbanek, S.: *rJava: Low-Level R to Java Interface* (2018), <https://CRAN.R-project.org/package=rJava>, r package version 0.9-10
21. Wijdicks, E.F., Klein, C.J.: Guillain-Barré syndrome. *Mayo Clinic Proceedings* 92(3), 467–479 (2017)
22. Witten, I.H., Frank, E., Hall, M.A., Pañ, C.: *Data Mining, Practical Machine Learning Tools and Techniques*. Morgan Kaufmann (2017)