Clasificación a partir de conjuntos de datos no equilibrados. Un marco para mejorar la aplicación de las estrategias de remuestreo

Resumen en español

TESIS DOCTORAL

PROGRAMA DE DOCTORADO EN INGENIERÍA INFORMÁTICA

Universidad de Salamanca

VNiVERSiDAD D SALAMANCA

Mohamed S. Kraiem

Salamanca, noviembre, 2020

UNIVERSIDAD DE SALAMANCA

TESIS DOCTORAL

Resumen en español

Clasificación a partir de conjuntos de datos no equilibrados. Un marco para mejorar la aplicación de las estrategias de remuestreo

---

Título original

Classification from imbalanced datasets. A framework for improving the application of sampling strategies

---

Author: Mohamed S. Kriem

Supervisor: María N. Moreno García

Salamanca, noviembre, 2020

Resumen

En los últimos años, el problema de la clasificación de datos no equilibrados se ha convertido en uno de los temas de investigación más candentes en el área del aprendizaje supervisado, donde encontrar una solución adecuada sigue siendo un desafío. La baja fiabilidad de los resultados de la clasificación a partir de datos desequilibrados se produce debido al sesgo del modelo predictivo hacia la clase mayoritaria, mientras que la clase minoritaria es casi ignorada o asumida como ruido por estar representada por muy pocas instancias. EL problema consiste en que los conjuntos de datos utilizados por los clasificadores suelen tener una distribución diferente de las instancias de cada clase. Esta situación, llamada clasificación desequilibrada de los conjuntos de datos, produce un bajo rendimiento predictivo para los ejemplos de la clase minoritaria. Por consiguiente, el modelo de predicción no suele ser válido, aunque la exactitud global del modelo pueda ser aceptable, ya que se obtiene principalmente de la correcta clasificación de los ejemplos de clase mayoritaria. Para resolver este problema se suelen utilizar algunas estrategias como el sobremuestreo y el submuestreo, que son procedimientos reconocidos cuyo objetivo es equilibrar el número de ejemplos de cada clase. Sin embargo, la eficiencia de esas estrategias se ve afectada por algunos factores como el solapamiento entre las clases, el tamaño del conjunto de datos, los ejemplos en el límite entre clases, el índice de desequilibrio, las características intrínsecas de los datos y el ruido en los datos, entre otros.

Esta investigación se divide en dos partes, la primera de las cuales comprende un estudio preliminar sobre el comportamiento de diferentes algoritmos de clasificación en contextos de datos no equilibrados, con diferentes índices de desequilibrio, antes y después de ser tratados con diferentes estrategias de remuestreo. El objetivo principal de estos experimentos iniciales es proporcionar una referencia que ayude a seleccionar el algoritmo con el mejor comportamiento para llevar a cabo el estudio posterior, que constituye la segunda y más importante parte de esta investigación. En esta parte se examinan diferentes factores relacionados con las características del conjunto de datos para determinar tanto el método de remuestreo más adecuado en función de las características del conjunto de datos, como las ventajas y los inconvenientes de las técnicas de remuestreo básicas y avanzadas. Los factores que se analizan en el estudio son el índice de desequilibrio, el solapamiento entre clases, los ejemplos en el límite entre clases, el tamaño reducido de la muestra, el número de instancias y el número de atributos. Se han utilizado diversas medidas de evaluación para contrastar los resultados de los diferentes modelos inducidos a partir de conjuntos de datos desequilibrados antes y después de procesarlos mediante estrategias de remuestreo básicas y avanzadas. Se han utilizado algunas métricas generales, como la precisión, y algunas medidas específicas para la clasificación de datos desequilibrados, como OP (*Optimized Precisión*) e IBA (Index of Balanced Accuracy). En esta parte se realizaron experimentos con conjuntos de datos de una amplia gama de características, los cuales fueron preprocesados con siete técnicas de remuestreo. El algoritmo de clasificación seleccionado para este estudio fue Random Forest, debido a que los experimentos anteriores demostraron su mejor comportamiento en contextos de datos desequilibrados en comparación con los otros algoritmos de aprendizaje supervisado.

Índice de contenidos

1. INTRODUCCION

La clasificación a partir de datos desequilibrados es actualmente un tema relevante para muchos investigadores debido a su importancia y a su presencia en muchos ámbitos de aplicación. Esta característica de los datos representa un importante inconveniente, conocido como el problema del desequilibrio de clases, y su solución es aún un desafío en el campo del aprendizaje supervisado. Muchos conjuntos de datos del mundo real presentan una distribución muy desequilibrada de las instancias de las clases. Así, una sola clase, conocida como la clase minoritaria, tiene muchos menos ejemplos que la otra clase (o clases), la clase (o clases) mayoritaria [1]. La mayoría de los clasificadores son muy sensibles a las distribuciones desiguales de las clases, proporcionando un resultado que se basa principalmente en la clase mayoritaria e ignora la clase minoritaria. En consecuencia, la precisión para esta última clase suele ser mucho menor que para la que tiene el mayor número de instancias, lo que indica un modelo de predicción poco fiable. La clase minoritaria es a menudo la más interesante en el área de aplicación, lo que agrava aún más el problema.

Son muchos los dominios de aplicación en los que es necesario clasificar conjuntos de datos desequilibrados para descubrir patrones y extraer información útil. Entre ellos se pueden incluir la detección de transacciones de cuentas bancarias o llamadas telefónicas fraudulentas, el diagnóstico biomédico, la clasificación de textos, la recuperación de información, el filtrado y la retención de estudiantes universitarios. Estas aplicaciones son ejemplos de varios ámbitos de aplicación que muestran la importancia del problema de la clasificación desequilibrada. Los clasificadores tradicionales inducidos a partir de estos conjuntos de datos suelen estar sesgados hacia la clase mayoritaria, por lo que no clasifican adecuadamente los ejemplos de las clases minoritarias, lo que limita su aplicabilidad [2].

Para hacer frente a esta situación, se han hecho varias propuestas tanto a nivel de datos como de algoritmos, aunque las técnicas de sobremuestreo y submuestreo son las estrategias más conocidas y utilizadas. Su objetivo es equilibrar el diferente número de ejemplos de cada clase. El sobremuestreo de conjuntos de datos implica la réplica del número de ejemplos de la clase minoritaria, mientras que el submuestreo es el proceso de eliminación de ejemplos de la clase mayoritaria. Estas estrategias se ven afectadas por algunos factores como las características intrínsecas de los datos, el solapamiento entre las clases, los datos con ruido, el tamaño del conjunto de datos y los ejemplos en el límite entre clases, entre otros [3].

Nuestro estudio aborda el problema de la clasificación de los datos desequilibrados analizando la eficacia de las estrategias de remuestreo y la influencia de los factores indicados anteriormente. Se divide en dos partes, un análisis preliminar del comportamiento de los algoritmos de clasificación conocidos en diferentes contextos de datos desequilibrados, y una segunda parte, centrada en el mejor algoritmo previamente identificado, en la que se realiza un

exhaustivo estudio experimental para alcanzar los objetivos perseguidos en esta investigación. Los resultados del estudio permiten determinar qué método de remuestreo es el más adecuado en función de las características del conjunto de datos, y conocer las ventajas e inconvenientes de las técnicas de remuestreo básicas y avanzadas.

En la primera parte hemos estudiado el comportamiento de los modelos inducidos por los algoritmos de aprendizaje automático comúnmente usados en una amplia gama de diferentes conjuntos de datos desequilibrados, cada uno de los cuales se procesa mediante diferentes estrategias de remuestreo. El estudio tiene por objeto identificar el algoritmo más apropiado para estos contextos. Esta parte comprende los experimentos realizados con más de 40 conjuntos de datos de diferentes tamaños y diferentes índices de desequilibrio que se procesaron mediante siete algoritmos de remuestreo.

El objetivo de la investigación desarrollada en la segunda parte es estudiar en profundidad el efecto de algunos factores relativos a las características de los conjuntos de datos sobre la fiabilidad de los modelos de clasificación obtenidos a partir de los conjuntos de datos originales y remuestreados, aplicando el mejor algoritmo de aprendizaje automático identificado en la primera parte del estudio. Para ello, se llevó a cabo un amplio estudio en el que se examinan estos factores para determinar el método de muestreo más adecuado según las características del conjunto de datos. Con el fin de lograr un análisis profundo de los resultados, se aplicaron varias medidas de evaluación para comparar los resultados de los modelos inducidos a partir de conjuntos de datos desequilibrados antes y después de su preprocesamiento con estrategias de remuestreo básicas y avanzadas. Se utilizaron medidas generales como la precisión, *recall*, medida F, media geométrica o AUC (*Area Under the ROC Curve*) y medidas específicas de los problemas de desequilibrio, como como OP (*Optimized Precisión*) [4] e IBA (Index of Balanced Accuracy) [5], entre otras. En esta parte de la investigación se utilizó *Random Forest* para inducir los modelos, dado que los resultados de la primera parte demostraron que este algoritmo muestra un mejor comportamiento en contextos de datos desequilibrados frente a otros algoritmos [6].

2. OBJETIVOS

El objetivo principal de esta tesis es abordar el problema de la clasificación de datos no equilibrados, concretamente su tratamiento mediante estrategias de remuestreo básicas y avanzadas. Como la mayoría de los métodos, estas técnicas tienen ventajas e inconvenientes que es necesario analizar para diferentes contextos y diversos conjuntos de datos, ya que su comportamiento depende de muchos factores.

El objetivo es realizar un estudio comparativo profundo sobre el comportamiento de los algoritmos de aprendizaje automático en contextos de datos desequilibrados y comprobar la eficacia de las estrategias de remuestreo. Los objetivos específicos que se pretenden alcanzar son los siguientes:

- Estudio del comportamiento de los clasificadores inducidos con varios algoritmos de aprendizaje automático a partir de conjuntos de datos desequilibrados.
- Estudio y análisis del efecto del índice de desequilibrio en la fiabilidad de los modelos de clasificación y la eficacia de las estrategias de remuestreo, ya que es uno de los factores más influyentes.
- Examen del tamaño, la complejidad, la dispersión, el grado de desequilibrio y otras características de los conjuntos de datos que pueden influir en el éxito de las estrategias de remuestreo.
- Análisis de las mejoras en el comportamiento de los clasificadores más utilizados cuando los conjuntos de datos desequilibrados originales se preprocesan mediante estrategias de remuestreo básicas y avanzadas.
- Obtención de resultados de evaluación a partir de una amplia gama de métricas, tanto las que se aplican generalmente en problemas de clasificación como las específicas para la clasificación desequilibrada.
- Análisis de las ventajas e inconvenientes de los métodos de remuestreo en función del procedimiento, las características analizadas de los datos y los algoritmos aplicados de aprendizaje automático.
- Propuesta y validación de un marco que sirva de apoyo para seleccionar la combinación más adecuada de los algoritmos de aprendizaje automático y las técnicas de muestreo para el problema que se ha de resolver.

3. ESTADO DEL ARTE

3.1. Estrategias de remuestreo

La forma más fácil de manejar el problema del desequilibrio de clases es la aplicación de los métodos básicos de remuestreo. Estos métodos incluyen el submuestreo aleatorio (RUS-*Random Under Sampling*), el sobremuestreo aleatorio (ROS-*Random Over Sampling*) y la combinación de ambos.

RUS [7] elimina ejemplos de la clase mayoritaria para equilibrar el conjunto de datos. La desventaja de este método es que ignora información potencialmente útil que puede ser importante para inducir a los clasificadores. Esta pérdida de información conduce a un menor rendimiento de los modelos. Además, RUS reduce el tamaño de la muestra, lo que contribuye aún más al empeoramiento del modelo. Este método es adecuado para conjuntos de datos con bajo desequilibrio y no es en absoluto apropiado para distribuciones altamente desequilibradas y con bajo número de instancias.

ROS [7] equilibra el conjunto de datos replicando los ejemplos de la clase minoritaria. La ventaja de este método es que no produce pérdida de información como en RUS. La desventaja es que conduce a un sobreajuste e introduce un coste computacional adicional si el índice de desequilibrio del conjunto de datos es alto. Además, la técnica ROS aumenta la proporción de solapamiento entre instancias de diferentes clases, debido a la duplicación de datos, ya que la replicación de las instancias hace que la similitud entre los atributos sea mayor. ROS es muy eficiente para los conjuntos de datos con un alto índice de desequilibrio [6].

Los métodos avanzados de remuestreo utilizan cierta inteligencia al eliminar o añadir ejemplos para crear distribuciones equilibradas de datos. Esto puede reducir las deficiencias de los métodos básicos descritas previamente.

Las técnicas de sobremuestreo aleatorio se han modificado para abordar algunos de los principales inconvenientes de las técnicas básicas, evolucionando hacia enfoques más avanzados. Uno de los más populares es el SMOTE (Synthetic Minority Over-Sampling Technique) [8], un método de sobremuestreo propuesto para abordar el problema del sobreajuste, haciendo más general el límite de decisión de la clase minoritaria. En este método, en lugar de replicar las instancias de la clase minoritaria, se generan nuevos ejemplos interpolando los ya existentes. SMOTE es un algoritmo especialmente eficiente para conjuntos de datos altamente desequilibrados y su rendimiento puede mejorarse combinándolo con técnicas de submuestreo [8].

Tomek Link (T-Link) [9] es otra técnica avanzada de muestreo propuesta para mejorar el método de clasificación del vecino más cercano mediante la eliminación del ruido y los ejemplos en el borde de las clases. Aunque el objetivo del método no es devolver datos equilibrados, puede

considerarse como un método de submuestreo guiado cuando se eliminan las instancias de formación de la clase mayoritaria. La base de esta estrategia es la siguiente:

1. Se consideran dos instancias de entrenamiento x e y de diferentes clases.
2. La distancia entre dichas instancias es $d(x, y)$.
3. El par (x, y) se llama *T-link* si no hay una instancia z que cumple la condición $d(x, z) < d(x, y)$ o $d(y, z) < d(x, y)$.
4. Si dos instancias cualesquiera son un *T-link*, o una de ellas es ruido o ambas instancias están en el límite de las clases.

La estrategia de selección unilateral (OSS- *One Sided Selection*) es una técnica de submuestreo propuesta por kubat y Matin [10], que elimina los casos de la clase mayoritaria que se consideran redundantes. El procedimiento es el siguiente:

1. D es el conjunto de datos original.
2. P es un subconjunto de D que contiene inicialmente todos los ejemplos de la clase minoritaria de D y un ejemplo x de la clase mayoritaria seleccionado al azar.
3. Usando P como conjunto de entrenamiento, la regla de los vecinos más cercanos (KNN) con k= 1 se usa para clasificar las instancias restantes de la clase mayoritaria.
4. Todos los ejemplos mal clasificados se trasladan al conjunto de entrenamiento P, que es más pequeño que D. Las instancias correctamente clasificadas de la clase mayoritaria se descartan ya que se consideran redundantes.
5. Finalmente, el *T-link* se utiliza para la limpieza de datos. Todos los ejemplos limítrofes y/o ruido de la clase mayoritaria se eliminan de P, mientras que todos los ejemplos de la clase minoritaria se conservan.

OSS está indicado para conjuntos de datos poco desequilibrados, pero puede utilizarse combinado con otros algoritmos para procesar conjuntos de datos con alto desequilibrio. En nuestro estudio se combina con SMOTE.

NCL (*Neighborhood Cleaning Rule*) [11] es una técnica de submuestreo que utiliza la regla de Wilson ENN (*Edited Nearest neighbor*) para identificar los datos de ruido y eliminar algunas instancias de entrenamiento de la clase mayoritaria. La regla ENN identifica los tres vecinos más cercanos (NN) de cada instancia de entrenamiento. Esta instancia se elimina si pertenece a la clase mayoritaria y al menos dos de sus tres NN pertenecen a una clase diferente. De manera similar, en la presente investigación, combinamos el método SMOTE con el método NCL.

CNN (*Condensed Nearest Neighbor*) es una estrategia propuesta inicialmente por Hart [12] como un método de reducción de datos utilizado para mejorar la eficiencia en la aplicación de la regla de decisión del vecino más cercano para los problemas de clasificación. Este método clasifica los ejemplos del conjunto de entrenamiento en tres tipos:

- Valores atípicos: puntos cuyos k vecinos más cercanos no pertenecen a la misma clase.
- Prototipos: el conjunto mínimo de puntos del conjunto de entrenamiento necesario para clasificar correctamente los puntos atípicos.
- Puntos absorbidos: puntos que se clasificarían correctamente a partir del conjunto de prototipos.

CNN identifica los puntos del prototipo del conjunto de datos original, que se utilizarán para clasificar las nuevas instancias. Los puntos absorbidos y los valores atípicos no se utilizan para la clasificación. El algoritmo funciona como se describe a continuación:

1. D denota el conjunto de datos original y E el conjunto condensado resultante que contiene los prototipos.
6. Se selecciona un punto arbitrario de D y se coloca en un conjunto original vacío E.
7. Los puntos restantes de D se clasifican por la regla NN utilizando E y los que se clasifican incorrectamente se añaden a E.
8. Este procedimiento se repite hasta que no se transfieren más puntos de datos de D a E.

Existen también estrategias híbridas en las que las técnicas de sobremuestreo y submuestreo se aplican de forma combinada para obtener mejores resultados. De esta forma, el conjunto de datos puede equilibrarse con pocas pérdidas de información, aunque, los inconvenientes de estas estrategias siguen presentes en este enfoque. Dos ejemplos de técnicas híbridas, que se han utilizado comúnmente incluyen SMOTE+Tomek o SMOTE+ENN [7], donde SMOTE [8] se utiliza para sobremuestrear la clase minoritaria, mientras que Tomek y ENN, respectivamente, se utilizan para submuestrear la clase mayoritaria.

3.2. Trabajos relacionados

El primer estudio importante realizado para abordar el problema del desequilibrio de clases se publicó en 2000. Japkowicz y Stephen [13] llevaron a cabo experimentos en 125 conjuntos de datos de diferentes tamaños, complejidad y distribución de clases utilizando redes neuronales. El estudio determinó que cuanto más complejo es el conjunto de datos, mayor es el grado de desequilibrio. La complejidad de los datos se midió mediante varias métricas, como la relación discriminante de Fisher. Varios estudios han puesto de relieve otras razones que explican los malos resultados de la clasificación a partir de datos no equilibrados.

Según [14], el comportamiento de los clasificadores depende de factores como el indice de desequilibrio entre la clase mayoritaria y la minoritaria, las propiedades del conjunto de datos, los algoritmos de clasificación y los métodos de remuestreo. La fiabilidad de los modelos de clasificación puede mejorarse en gran medida cuando se reduce el desequilibrio entre la clase mayoritaria y la minoritaria utilizando los métodos de remuestreo apropiados.

En [13], se discute el efecto de los métodos básicos de remuestreo en los resultados del árbol de decisión C5.0. Los autores llegaron a la conclusión de que el índice de desequilibrio es un factor importante cuyo valor es propocional a la tasa de error del clasificador, aunque esta tasa disminuye a medida que aumenta el tamaño del conjunto de entrenamiento.

López y otros [3] han realizado una revisión de los métodos propuestos para tratar el problema de la clasificación de datos desequilibrados y han propuesto una taxonomía para clasificarlos. El objetivo del trabajo es analizar los métodos de preprocesamiento, incluyendo las estrategias de remuestreo, y comparar su eficacia como paso previo a la clasificación. En el estudio se analizan algunas características intrínsecas de los datos, en el que los autores concluyen que cuestiones como la falta de densidad, el pequeño tamaño de la muestra y el solapamiento de clases tienen una mayor influencia en el rendimiento de los clasificadores que el índice de desequilibrio. Sin

embargo, estas conclusiones no pueden generalizarse, ya que los conjuntos de datos utilizados en el estudio tienen un bajo coeficiente de desequilibrio y la mayoría de ellos tienen un número reducido de atributos e instancias.

El manejo del problema de los conjuntos de datos no equilibrados y la superposición entre clases es el objetivo del trabajo presentado en [15]. En este estudio, los autores proporcionan algunos métodos eficaces para resolver ambos inconvenientes. En este contexto, los autores sugirieron el uso de redes neuronales y métodos de selección de instancias para mejorar el rendimiento.

El impacto del índice de desequilibrio para un tipo específico de clasificadores también se ha analizado en otros estudios cuyo propósito principal es ayudar a seleccionar el mejor método de remuestreo en relación con los valores de ese factor [16]. Otros trabajos centrados en algoritmos de clasificación específicos se dedican a proponer enfoques de muestreo para mejorar su rendimiento. En [17] un procedimiento híbrido que combina tanto el sobremuestreo como el submuestreo con un *ensemble* de SVM logra mejorar la fiabilidad de ese clasificador.

Algunos trabajos de la bibliografía se centran en propuestas para mejorar las estrategias de remuestreo tradicionales en ámbitos de aplicación específicos. En [18], SVM (*Support Vector Machines*) se utiliza conjuntamente con SMOTE para el preprocesamiento de datos desequilibrados relacionados con clientes de caravanas. Cateni y otros [19] propusieron un nuevo método de remuestreo combinando técnicas de sobremuestreo y submuestreo. En este estudio los autores tratan de reducir al mínimo los problemas asociados a estas dos técnicas. Otra investigación encaminada a mejorar los métodos de remuestreo existentes es la realizada por Sáez y otros [20]. Estos autores amplían SMOTE añadiendo un filtro de ruido basado en un conjunto denominado Filtro de Partición Iterativo (IPF) a fin de abordar los problemas en la clasificación producidos por ejemplos con ruido. Los resultados de esta extensión, denominada SMOTE-IPF, son mejores que los obtenidos por el SMOTE básico y otros métodos basados en SMOTE.

Por otra parte, se han realizado numerosos estudios comparativos en los que se evalúa la efectividad de diferentes estrategias de remuestreo [21-23] tanto en un dominio de aplicación particular como con datos de diferentes dominios. Sin embargo, en estos estudios no se analiza el efecto de las características intrínsecas de los datos.

4. PLANTEAMIENTO DEL PROBLEMA

En este capítulo se examinan algunos aspectos relativos a las estrategias de remuestreo a fin de contextualizar el problema que se va a abordar. El propósito de este trabajo es mostrar la eficacia y factibilidad de estos métodos en la clasificación datos no equilibrados, por lo que en este capítulo se señalan sus principales ventajas e inconvenientes, así como sus efectos conocidos en función de las características de los datos. Además, se describen las métricas más adecuadas para la evaluación de los modelos de clasificación creados a partir de datos con desequilibrio.

4.1. Ventajas e inconvenientes de las estrategias de remuestreo

El remuestreo es un enfoque común para manejar el problema del desequilibrio de clases. Aunque estas técnicas mejoran los resultados de la clasificación, no están exentas de problemas.

Las ventajas de RUS son su fácil aplicación y el hecho de que puede ayudar a mejorar los problemas de tiempo de ejecución y de almacenamiento al reducir el número de instancias de del conjunto de entrenamiento cuando este es muy grande. Sin embargo, los inconvenientes de esta estrategia son la pérdida de información importante, especialmente cuando los índices de desequilibrio son altos. Además, puede causar el problema de tamaño de muestra pequeño, lo cual conduce a una menor fiabilidad del modelo.

Por el contrario, en ROS los ejemplos de la clase minoritaria se multiplican hasta que se alcanza una distribución más equilibrada. Este método no conduce a la pérdida de información y es fácil de aplicar, pero puede causar un sobreajuste y un aumento del tamaño del conjunto de datos, especialmente en el caso de alto desequilibrio, lo que da lugar a un mayor tiempo para entrenar el modelo.

SMOTE es una técnica avanzada de remuestreo cuya principal ventaja es mitigar el problema del sobreajuste causado por RUS. SMOTE es eficiente cuando se combina con estrategias de submuestreo, pero no es muy eficaz cuando se aplica a conjuntos de datos de alta dimensionalidad. Además, produce un aumento de ejemplos en el límite de las clases.

Las técnicas avanzadas de submuestreo como OSS, Tomek Link, NCL y CNN no se utilizan para devolver conjuntos de datos equilibrados sino para eliminar ruido en los datos, por lo que no causan pérdida de información. Sin embargo, tienen algunos inconvenientes que dependen de la técnica particular. Cuando se utiliza Tomek Link, muchas muestras pueden ser eliminadas si el límite no está claro, de manera que en ese caso, CNN puede eliminar información importante. Estos métodos son muy útiles cuando se combinan con técnicas de sobremuestreo como SMOTE. La combinación de SMOTE y los métodos avanzados de submuestreo reduce los efectos de los ejemplos límite y el problema de solapamiento de clases.

En este trabajo se han analizado las estrategias anteriores tanto de forma aislada como combinadas.

4.2. Influencia de las características de los datos

La clasificación a partir de datos desequilibrados es un problema complejo, ya que se ve afectada por muchos factores. Detectar los factores relativos a las características de los datos y descubrir su influencia en los resultados de la clasificación y en la efectividad de las estrategias de remuestreo es de gran interés, debido a la presencia de este problema en muchas áreas.

El índice de desequilibrio se define como la relación entre el número de ejemplos de la clase mayoritaria y el número de ejemplos de la clase minoritaria [24]. El desequilibrio se produce en muchos conjuntos de datos del mundo real en múltiples dominios de aplicación. Un índice de desequilibrio bajo siempre da lugar a mejores resultados de los modelos de clasificación.

El solapamiento entre clases es otro factor que afecta al rendimiento del modelo, especialmente dada la distribución desequilibrada de las clases. El solapamiento dificulta a los clasificadores la identificación de los límites de una buena decisión entre las clases, lo que da lugar a una peor clasificación. Este problema se complica con la presencia de distribuciones de datos desequilibradas debido a la necesidad de utilizar estrategias para equilibrar los datos. La aplicación de métodos de sobremuestreo, supone la repetición de instancias de la clase minoritaria, especialmente cuando el porcentaje de desequilibrio es elevado. En este caso, la similitud entre los ejemplos lleva a un aumento del solapamiento entre las clases.

El problema de los ejemplos en el límite entre clases también repercute en la reducción de la fiabilidad del clasificador [25]. Este problema está relacionado con el ruido en los datos y su influencia en el comportamiento del modelo es mayor que el del solapamiento. Cuando se aplican técnicas de sobremuestreo, como ROS o SMOTE, el porcentaje de ejemplos límite se incrementa al aumentar el número de instancias.

Otro problema que surge en la clasificación es el pequeño tamaño de la muestra referido al conjunto de entrenamiento. Este factor está relacionado con la falta de información o la baja densidad. En este caso, el algoritmo no tiene suficiente información para construir y validar el modelo. La situación se complica en las distribuciones no equilibradas con la aplicación de métodos de submuestreo. La mayoría de los trabajos en la literatura confirman que cuanto mayor sea el número de instancias, menor será el índice de error en la clasificación.

Estas y otras características serán analizadas en el estudio experimental realizado en este trabajo.

4.3. Métricas para la evaluación de los clasificadores

La aplicación de medidas de evaluación inadecuadas para evaluar un modelo de clasificación creado a partir de un conjunto de datos desequilibrado puede dar lugar a una interpretación errónea de los resultados. Si sólo se utiliza la exactitud para medir la calidad del modelo, podemos obtener buenos resultados incluso con malos clasificadores. Esto ocurre cuando el porcentaje de instancias correctamente clasificadas es alto debido a la clasificación correcta de un porcentaje muy alto de instancias de la clase mayoritaria, aunque la tasa de error en la

clasificación de instancias de la clase minoritaria sea muy alta. De ahí que se requieren otras medidas de evaluación alternativas.

Además de las métricas tradicionales de exactitud, precisión, *recall*, *F-measure*, etc., existen otras métricas más apropiadas para evaluar los clasificadores cuando existe desequilibrio de datos. Entre estas se encuentra la métrica G-mean que evalúa la precisión de cada clase tomando la media geométrica de la tasa de verdaderos positivos (TP rate) y la tasa de verdaderos negativos (TN rate). Un valor bajo de G-mean indica un clasificador que está altamente sesgado hacia una de las clases.

$$G - mean = \sqrt{TP\ rate\ \times TN\ rate}$$

Otra de las técnicas de validación más populares para los problemas de clasificación de datos desequilibrados es el área bajo la curva ROC (AUC).

Las métricas anteriores no tienen en cuenta el comportamiento del clasificador con la clase minoritaria. Para solventar esta deficiencia, Ranawana y Palade [4] propusieron una nueva métrica llamada *Optimized Precision* (OP).

$$OP = Precision - |TN\ rate - TP\ rate|/|TN\ rate + TP\ rate|$$

El mayor valor de OP se consigue cuando la precisión es alta y las dos tasas (TN rate y TP rate) son similares.

Recientemente, García et al [5] han propuesto una nueva medida llamada *Generalized Index of Balanced Accuracy* (IBA) que puede utilizarse y definirse para cualquier otra métrica de evaluación m:

$$IBA(m) = (1 + \alpha\ \ Dom)\ m$$

Donde, Dom es un índice llamado dominancia que se define como:

$$Dom = TP\ rate - TN\ rate$$

Los valores de Dom están dentro del rango [-1, +1]. El mejor valor de Dom se produce cuando la tasa de verdaderos positivos y la tasa de verdaderos negativos son aproximadamente iguales, lo que produce un valor de Dom cercano a cero. En general, este valor se pondera con el parámetro $\alpha \geq 0$ para ajustar su influencia en función de la métrica utilizada (m).

5. Estudio experimental

El estudio experimental llevado a cabo en esta tesis se divide en dos partes. La primera parte tiene por objeto determinar el algoritmo con mejor comportamiento en la clasificación de diferentes conjuntos de datos no equilibrados tras la aplicación de algunas estrategias de remuestreo. En la segunda parte, este algoritmo se aplica a una amplia variedad de conjuntos de datos con distintos índices de desequilibrio con el fin de analizar la efectividad de las técnicas de remuestreo más populares en función de las características de los datos sobre los que se aplican. El objetivo final es proporcionar un marco general que permita seleccionar la estrategia más adecuada para cada conjunto de datos según sus características.

Para alcanzar el primer objetivo, se analizaron los clasificadores inducidos con seis populares algoritmos de aprendizaje automático: *K-nearest neighbor* (KNN), *Naïve Bayes* (NB), *Support Vector Machines* (SVM) y tres algoritmos de la familia de los árboles de decisión (*Random Forest*-RF, J48 y C5.0). Estos se aplicaron a múltiples conjuntos de datos preprocesados mediante siete métodos de remuestreo: RUS, ROS, SMOTE, SMOTE+OOS, SMOTE+Tomek link, SMOTE+CNN y SMOTE+NCL.

En la segunda parte del estudio se realizaron varios experimentos en los que el algoritmo seleccionado en la primera parte se aplicó a conjuntos de datos que abarcan una amplia variedad de características, ya que, como se ha comentado en secciones anteriores, la eficacia de las estrategias de remuestreo se ve afectada por algunos factores asociados a las propiedades intrínsecas de los datos. En esta parte se analizan las repercusiones del índice de desequilibrio, el número de instancias, el número de atributos, el porcentaje de ejemplos en el límite y el solapamiento entre clases, mediante un amplio estudio comparativo en el que se examinan los resultados de la clasificación para determinar qué método de remuestreo es más adecuado en relación con las características del conjunto de datos. A fin de abarcar una amplia gama de condiciones, los experimentos se realizaron sobre 40 conjuntos de datos de diferentes ámbitos de aplicación. Esos conjuntos de datos abarcan un gran intervalo de valores de los factores en estudio.

En todos los experimentos realizados para evaluar los resultados de los clasificadores, los conjuntos de entrenamiento y de prueba se generaron utilizando validación cruzada. Un aspecto importante de la validación es el hecho de que las estrategias de remuestreo se aplicaron sólo a las particiones de datos de los conjuntos de entrenamiento en cada iteración, mientras que las particiones de los conjunto de prueba no se remuestrearon. El objetivo es asegurar que no haya problemas de sobreajuste y que los clasificadores que den buenos resultados puedan aplicarse a ejemplos reales diferentes del conjunto de entrenamiento. Cuando los conjuntos de pruebas se forman a partir de los ejemplos remuestreados, los resultados suelen ser mejores, pero es probable que estos resultados no se reproduzcan cuando se utilicen los clasificadores en un contexto real.

Con el fin de lograr un análisis profundo de la fiabilidad de los clasificadores y la efectividad de las estrategias de remuestreo, se han aplicado ocho métricas de evaluación para comparar los resultados de los modelos, los cuales fueron inducidos a partir de conjuntos de datos desequilibrados antes y después de preprocesarlos mediante las siete estrategias de muestreo. Las métricas utilizadas en el estudio fueron las clásicas (AUC, precisión, *recall*, *G-mean* y *F-measure*) y las nuevas medidas definidas específicamente para los problemas de desequilibrio (OP e IBA), descritas en el capítulo anterior.

Inicialmente, los valores de las métricas se analizaron inicialmente de forma individual para cada una de las características de los datos a fin de descubrir su impacto en la eficacia de las técnicas de remuestreo y posteriormente de forma global mediante la aplicación de test de significación estadística. Finalmente, se indujeron modelos de reglas de asociación que permiten la selección automática de la estrategia de remuestreo más adecuada para cualquier conjunto de datos, sin importar a qué campo pertenezca, ya que la adecuación se basa únicamente en las propiedades de los datos. La gran diversidad de los conjuntos de datos y el volumen de los resultados utilizados para generar las reglas de asociación proporcionan a estos modelos una gran utilidad para la adopción de decisiones en una amplia variedad de condiciones.

Los conjuntos de datos fueron extraídos de diferentes fuentes:

- 22 conjuntos de datos del depósito de datos de KEEL.
- 2 conjuntos de datos del repositorio de datos UCI.
- 15 conjuntos de datos creados utilizando la función *make_classification* de la biblioteca Scikit-learn de Python.
- 1 conjunto de datos sobre infecciones de pacientes hospitalizados en la unidad de cuidados intensivos del Hospital Universitario de Salamanca, España.

6. Resultados

El estudio experimental descrito en el capítulo anterior proporcionó un volumen elevado de resultados de las métricas de validación aplicadas sobre todos los clasificadores, ya que estos se aplicaron sobre todos los conjuntos de datos, tanto originales como preprocesados con cada una de las siete estrategias de remuestreo. En este capítulo se presenta una muestra de estos resultados después de ser sometidos a un proceso de síntesis y agregación, además de a la aplicación de test de significación estadística con el fin de hacer un mejor análisis de los mismos.

El análisis preliminar correspondiente a la primera parte del estudio proporcionó los resultados de los seis algoritmos de aprendizaje referidos anteriormente con el objetivo de seleccionar el algoritmo que se utilizará en la segunda parte del estudio. Se utilizaron las métricas de AUC, precisión y *recall* para evaluar la fiabilidad de los clasificadores. Las figuras siguientes muestran el valor medio sobre todos los conjuntos de datos de cada una de estas tres métricas en los experimentos realizados.

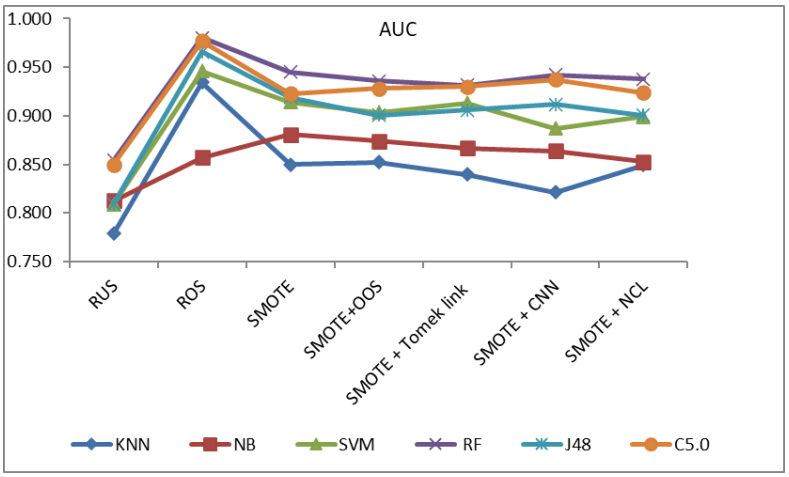

Fig. 1. AUC media de los clasificadores obtenidos de los datos remuestreados con diferentes técnicas.

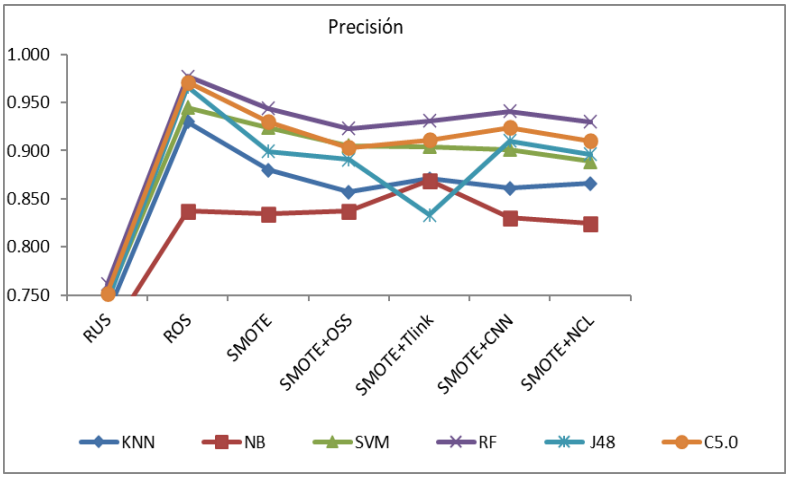

Fig. 2. Precisión media de los clasificadores obtenidos de los datos remuestreados con diferentes técnicas.

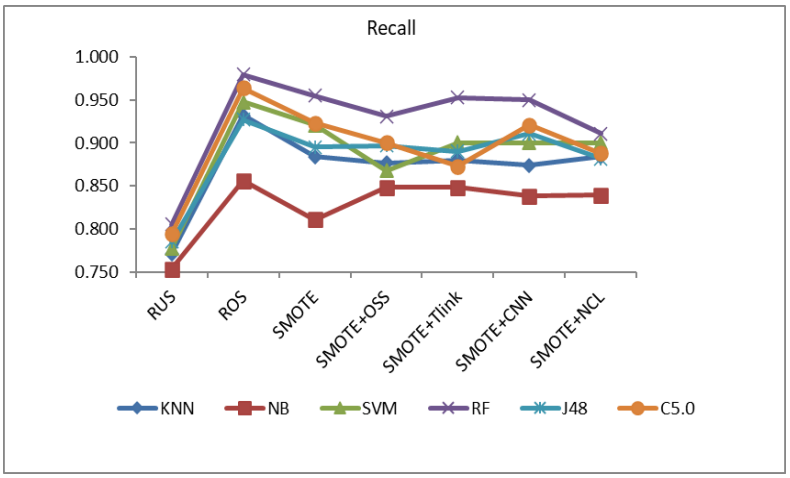

Fig. 3. *Recall* medio de los clasificadores obtenidos de los datos remuestreados con diferentes técnicas.

Adicionalmente, se llevaron a cabo pruebas de significación estadísticas no paramétricas y post hoc para realizar múltiples comparaciones con el fin de detectar diferencias significativas entre los resultados procedentes de los clasificadores obtenidos de múltiples conjuntos de datos. En este estudio, realizamos la prueba de Friedman ya que es un procedimiento popular y poderoso que permite detectar diferencias significativas para comparaciones múltiples.

Tanto los gráficos mostrados en las figuras anteriores como las pruebas estadísticas demostraron que el mejor algoritmo de clasificación con todas las estrategias de remuestreo es *Random Forest* (RF).

Una vez seleccionado *Random Forest* como el algoritmo que proporcionó los mejores resultados, este fue aplicado en la segunda parte del estudio en combinación con las estrategias de remuestreo para evaluar la eficacia de estas últimas en cada uno de los conjuntos de datos. Los conjuntos de datos originales se trataron con el fin de reducir su coeficiente de desequilibrio mediante las siguientes técnicas: RUS, ROS, SMOTE y SMOTE combinados con OSS, CNN, ENN y TL (T-Link). El porcentaje utilizado para el sobre-muestreo fue del 500% y para el sub-muestreo (RUS) del 50%. Como ya se ha comentado, las métricas utilizadas fueron exactitud, precisión,

recall, *F-measure*, AUC, *G-mean*, OP e IBA. Los resultados de las métricas se analizaron tanto de forma individual como globalmente, realizando también pruebas de significación estadística para comparar los resultados. Por último, se elaboraron modelos de reglas de asociación para determinar el método de remuestreo más adecuado según las características del conjunto de datos.

Para el examen individual de las métricas frente a cada característica estudiada, se representó el valor de las mismas para todos los conjuntos de datos en su forma original y remuestreados con cada una de las estrategias de remuestreo indicadas previamente. Se evaluaron 5 características de los datos (índice de desequilibrio, número de instancias, número de atributos, porcentaje de ejemplos en el límite y solapamiento entre clases) con 8 métricas diferentes, por lo que se elaboraron un total de 40 gráficos similares al de la figura 4. Adicionalmente, tanto para cada métrica como globalmente se hicieron los test estadísticos indicados previamente y los resultados post-hoc se mostraron mediante gráficos CD (*Critical Distance*) y gráficos de algoritmos (figura 5).

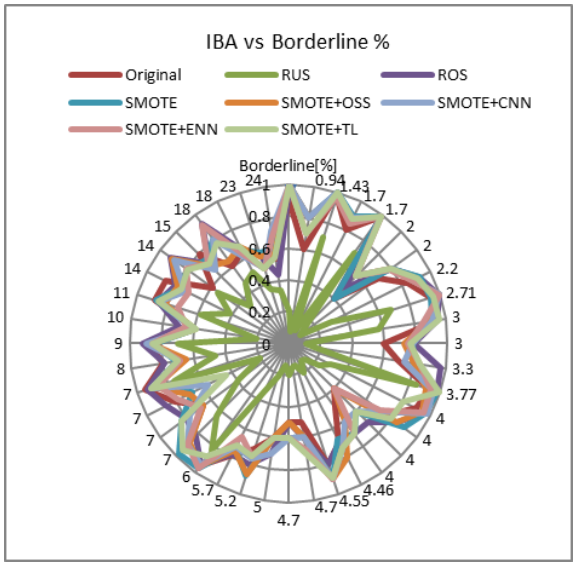

Fig. 4. Influencia en el valor de IBA del porcentaje de ejemplos en el límite de las clases (*borderline* %) para todas las estrategias de remuestreo.

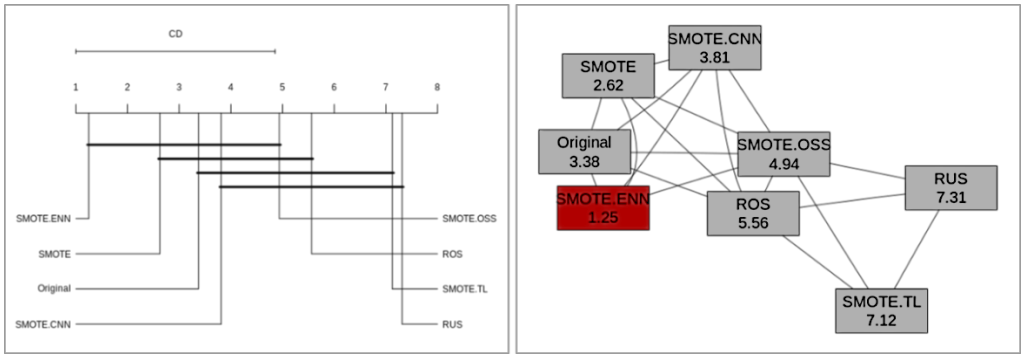

Fig. 5. Gráfico CD (izquierda) y gráfico de algoritmos (derecha) obtenidas por comparación de todas las métricas obtenidas para uno de los conjuntos de datos

Tras el análisis cualitativo de cada métrica de calidad y tras encontrar el mejor método de remuestreo para cada conjunto de datos, el siguiente paso fue realizar un análisis cuantitativo para obtener un conocimiento más profundo del impacto de las propiedades de los datos en las estrategias de remuestreo. Para ello, recurrimos a la inducción de reglas de asociación que relacionan las características de cada conjunto de datos con el método que proporcionó los mejores resultados tanto en lo que respecta a las métricas individuales como a las globales. Para ello, cada conjunto de datos con su información de propiedades se etiquetó con el método de remuestreo que le dio el mejor resultado, y luego se aplicó el conocido algoritmo Apriori para la extracción de reglas de asociación. Se utilizaron las medidas de calidad de confianza, *lift*, *leverage* y convicción para evaluar la calidad de las reglas de asociación.

La confianza se define como la proporción de ejemplos que contienen el consecuente y el antecedente de la regla en relación con la proporción de ejemplos que contienen el antecedente. Toma valores entre 0 y 1. *Lift* indica la probabilidad de la regla en relación con la probabilidad conjunta del antecedente y el consecuente si fueran independientes. Las reglas son válidas si los valores de lift son mayores que 1. La métrica *leverage* mide la diferencia entre la probabilidad de la regla y la probabilidad conjunta del antecedente y el consecuente si fueran independientes. Toma valores en el intervalo [-1, 1]. Convicción evalúa el grado en que el término precedente influye en la ocurrencia del término consecuente de una regla de asociación. Toma valores de 0 a infinito y la regla sólo es aceptable para valores superiores a 1.

El primer modelo de asociación se generó para métricas individuales. Teniendo en cuenta que las métricas más específicas para evaluar la clasificación de los datos no equilibrados son OP e IBA, se recurrió a la inducción de reglas de asociación que relacionan las características de cada conjunto de datos con el método que proporcionó el mejor valor de la métrica elegida, que en este caso fue IBA, ya que OP dio resultados muy similares. Para poder aplicar el algoritmo de Apriori, los atributos se discretizaron en 5 intervalos de igual tamaño. En la siguiente tabla se presentan algunas de las mejores reglas obtenidas y los correspondientes indicadores de calidad: Confianza (conf.), *lift*, *leverage* (lev.) y convicción (conv.).

Tabla 1. Evaluación mediante reglas de asociación del mejor método en cuanto a los valores de IBA

IR	#Inst.	#Attrib.	BL%	OVL%	Best IBA	Conf.	Lift	Lev.	Conv.
(-∞-48]	-	-	(6.75-12.56]	(78.8-∞)	ROS	1.00	9.00	0.03	1.78
(-∞-48]	-	-	(12.56-18.38]	(44.4-61.6]	Original	1.00	6.75	0.03	1.70
(48-91]	-	-	(-∞-6.75]	(78.8-∞)	ROS	1.00	1.64	0.01	0.78
(48-91]	-	-	(-∞-6.75]	(61.6-78.8]	SMOTE	1.00	1.64	0.01	0.78
-	-	-	(-∞-6.75]	(61.6-78.8]	SMOTE+CNN	1.00	1.64	0.01	0.78
-	-	(-∞-22]	(-∞-6.75]	-	SMOTE+OSS	1.00	1.29	0.02	1.11
-	(-∞-3154]	(-∞-22]	-	(61.6-78.8]	SMOTE	1.00	1.29	0.02	1.11
-	-	(-∞-22]	(-∞-6.75]	(61.6-78.8]	SMOTE	1.00	1.29	0.02	1.11
(-∞-48]	-	(-∞-22]	-	-	SMOTE+TL	1.00	1.29	0.02	0.89

El modelo completo de reglas de asociación nos permitirá seleccionar automáticamente la mejor estrategia de remuestreo para cada conjunto de datos. Examinando las reglas que se muestran en la tabla ya se pueden sacar algunas conclusiones. La primera regla significa que si los valores

de IR están entre - ∞ y 48, BL% entre -∞ y 6,75 y OVL% entre 78,8 y ∞, entonces el ROS tiene el mejor valor de IBA independientemente del número de instancias y el número de atributos. La segunda regla significa que si los valores de IR tienen valores entre - ∞ y 48, BL% entre 6,75 y 18,38 y OVL% entre 44,4 y 61,6, entonces el conjunto de datos original tiene el mejor valor de la IBA independientemente del número de instancias y el número de atributos. La forma de interpretar el resto de las reglas es la misma que en las dos primeras que hemos dado como ejemplo

Otras conclusiones más generales derivadas del análisis de las normas son las siguientes. El mejor valor de IBA viene dado por el conjunto de datos originales, sin remuestreo, cuando IR es muy bajo, independientemente de las demás propiedades del conjunto de datos, excepto en el caso de que el solapamiento entre clases sea muy alta, entonces la ROS es la mejor estrategia. También puede observarse que SMOTE utilizada solo o combinado con CNN, OOS o TL es la estrategia que obtiene mejores resultados en la mayoría de las reglas. Sin embargo, SMOTE+ENN no alcanza el mejor valor de IBA bajo ninguna circunstancia. Las reglas de asociación también confirman la observación de que RUS es la peor estrategia, ya que no aparece en ninguna de las reglas.

Un modelo de reglas similar se obtuvo para el análisis global con los resultados obtenidos por los test de significación estadística aplicados sobre todas las métricas.

7. Conclusiones

El objetivo de la investigación presentada en este trabajo de tesis ha sido proporcionar una guía útil para seleccionar la estrategia de remuestreo más adecuada en función de las características de los datos cuando se aplican algoritmos de clasificación a conjuntos de datos no equilibrados. Como se ha mencionado en capítulos anteriores, inicialmente se realizó un estudio preliminar consistente en una comparación empírica exhaustiva de la eficacia de diferentes estrategias de remuestreo básicas y avanzadas, teniendo en cuenta diferentes clasificadores. Se utilizaron 24 conjuntos de datos con una amplia gama de índices de desequilibrio, desde 1,82% a 129,44%, 6 clasificadores ampliamente utilizados y 7 estrategias de remuestreo. En la evaluación se utilizaron 6 métricas diferentes y se aplicaron test de significación estadística para confirmar la validez del análisis comparativo. Los resultados experimentales demostraron que los clasificadores de la familia de los árboles de decisión, especialmente *Random Forest*, presentan mejor comportamiento que los otros clasificadores en la mayoría de los casos. Este hallazgo nos permitió tomar la decisión de utilizar el algoritmo *Random Forest* en la segunda parte del estudio.

El principal propósito de este trabajo se alcanza en esta última parte de la investigación cuyos resultados en forma de modelos de reglas de asociación proporcionan indicaciones para ayudar a elegir la estrategia de remuestreo adecuada para cada problema particular. Esta parte incluye un amplio estudio experimental en el que se utilizaron 40 conjuntos de datos de muy diversos dominios de aplicación para inducir modelos de clasificación tanto en su forma original como remuestreados mediante 7 métodos de remuestreo diferentes.

Dado que Random Forest fue el algoritmo que presentó el mejor comportamiento en el estudio preliminar, todos los experimentos fueron realizados con este método, ya que la inclusión de más métodos daría lugar a una extensión excesiva de la investigación. Además, el comportamiento respecto a las mejoras logradas con los métodos de remuestreo fue muy similar para la mayoría de los algoritmos de clasificación probados. Los resultados se evaluaron a través de un total de ocho métricas, tanto clásicas como nuevas métricas diseñadas para el problema específico de los datos desequilibrados.

Un primer análisis de los resultados de estas métricas y de las pruebas estadísticas realizadas reveló algunos comportamientos que se utilizaron como entrada para crear un modelo de reglas de asociación que nos permitiera sacar conclusiones más profundas. El modelo de reglas de asociación nos permite seleccionar automáticamente la mejor estrategia de remuestreo para cada conjunto de datos. Sin embargo, se obtienen algunas conclusiones generales derivadas del análisis de las reglas. Las reglas muestran que el mejor valor de la IBA viene dado por el conjunto de datos original, sin remuestreo, cuando el índice de desequilibrio es muy bajo, independientemente de las demás propiedades del conjunto de datos, excepto en el caso de que la superposición entre clases sea muy alta, entonces la mejor estrategia es ROS. También puede observarse que SMOTE utilizado solo o combinado con CNN, OOS o TL es la estrategia que

obtiene mejores resultados en la mayoría de las reglas. Sin embargo, SMOTE+ENN no alcanza el mejor valor de la IBA bajo ninguna circunstancia. Las reglas de la asociación también confirman la observación de que RUS es la peor estrategia ya que no aparece en ninguna de las reglas. La mayoría de estas conclusiones pueden extenderse al modelo de reglas de asociación generado a partir de las pruebas estadísticas realizadas sobre los resultados de todas las métricas.

Por último, se estudió la forma en que cambian las características de los conjuntos de datos cuando se aplican las técnicas de remuestreo y el efecto que esto puede tener en el resultado de los clasificadores. Se observó que la tasa de aumento y disminución del valor de la tasa de desequilibrio desempeña un papel importante en el aumento y disminución del tamaño de los datos cuando se aplican estrategias de remuestreo, en las que el tamaño de los datos en el caso de un alto desequilibrio aumenta cuando se utilizan métodos de remuestreo como SMOTE o ROS. El aumento del tamaño del número de instancias o del número de atributos incrementa el porcentaje del grado de ejemplos límite y el ruido, lo que repercute negativamente en el rendimiento de los modelos de clasificación.

Referencias

1. He, H., García, E.A. Learning from imbalanced data. IEEE Trans. Knowl. Data Engine 21(9), pp. 1263–1284 (2009).
2. Breiman, L. Random Forest, Machine Learning, 45, pp. 5-32, (2001).
3. Lopez, V., Fernandez, A., Garcia, S., Palade, V., Herrera, F. An insight into classification with Imbalanced data: Empirical results and Current Trends on Using data intrinsic characteristics, Information Sciences, 250, pp. 113-141 (2013).
4. Ranawana, R., Palade, V. Optimized precision - a new measure for classifier performance evaluation. In: Pro-ceeding of the IEEE Congress on Computational Intelligence, Vancouver, Canada, pp. 2245–2261 (2006).
5. García, V., Mollineda, R.A., Sánchez, J.S. Index of balanced accuracy: a performance measure for skewed class distributions. In: Araujo, H., Mendonça, A.M., Pinho, A.J., Torres, M.I. (eds.) IbPRIA 2009. LNCS, vol. 5524, pp. 441–448. Springer, Heidelberg (2009).
6. Kraiem. M. S, Moreno.M.N Effectiveness of Basic and Advanced Sampling Strategies on the Classification of Imbalanced Data. A Comparative Study Using Classical and Novel Metrics, 12th International Conference on Hybrid Artificial Intelligent Systems, HAIS 2017, LNCS, volume 10334, pp. 233-245 (2017).
7. Batista, G.E.A.P.A, Prati, R.C, Monard, M.C. A study of the behavior of several methods for balancing machine learning training data, SIGKDD Explor Newslett, 6(1),20-29, (2004).
8. Chawla, N.V., Bowyer, K., Hall, L., Kebelmeyer, W.P. SMOTE: synthetic minority over sampling technique. J. Artif. Intell. Res. 16, pp. 321-357 (2002).
9. Tomek, I. A generalization of the K-NN rule, IEEE Trans. SMC 6, pp. 121–126 (1976).
10. Kubat, M., Matwin, S. Addressing the curse of imbalanced training sets: one side selection. In: Fisher, D.H. (ed.) ICML, pp. 179–186, Morgan Kaufmann, San Francisco (1979).
11. Laurikkala, J. Improving identification of difficult small classes by balancing class distribution, in: Quaglini, S., Barahona, P., Andreassen, S. (eds.) AIME 2001. LNCS, vol. 2101, pp. 63–66. Springer, Heidelberg (2001).
12. Hand, B.J., Batchelor, B.G. Experiments on the edited condensed nearest neighbor rule. Inf. Sci. 14(3), pp. 171-180 (1978).
13. Japkowicz N., Stephen S. The class imbalance problem: a systematic study, Intelligent Data Analysis 6(5), pp. 429-449 (2002).
14. Hulse, J.V., Khoshgoftaar, T.M., Naplolitano, A. Experimental perspectives on learning from imbalanced data, in: Proceedings of the 24th International Conference on Machine Learning, Corvalis, Oregon, pp.935-942 (2007).
15. Lavanya, S., Polaniswam, S., Sudha, S. Efficient methods to solve class imbalance and class overlap, International journal of science Engineering and technology research, 3, pp. 1-5 (2014).

16. Loyola-Gonzalez, O., Martinez-Trinidad, F.J., Carrasco-Ochoa, J.A. Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases, Neurocomputing, 175, pp. 935–947 (2016).
17. Liu, Y., Yu, X., Huang J. X., An, A. Combining integrated sampling with SVM ensembles for learning from imbalanced datasets, Information Processing & Management, Volume 47, Issue 4, July 2011, pp. 617-631 (2011).
18. Farquad, M.A.H, Indranil, B.O.S. Preprocessing unbalance data using support vector machine, Decision Support Systems, 53 (2012), pp.226-233 (2012).
19. Cateni, S., Colla, V., Vannucci, M. A method for resampling imbalanced datasets in binary classification tasks for real world problems, Neurocomputing, 135:32-41,(2014).
20. Saez, J.A., Luengo, J., Stefanoroski, J., Herrera, F. SMOTE-IBF: Addressing the noisy and borderline examples problem in imbalanced classification by resampling method with filtering. Information Sciences, 291,184-203,(2015).
21. More, A. Survey of resampling techniques for improving classification performance in imbalanced datasets, arXiv:1608,06048V1 [Stat.AP]:1-7, (2016).
22. Adnan, A. et al., Comparing oversampling techniques to handle the class imbalance problem: A Customer Churun prediction Case Study, IEEE Access, 10(1190), 1-19, (2016).
23. Barandela. R, Valdovinos, R.M, Sanchez, J.S, Ferri, F.J, The imbalance training sample problem: under or over sampling, in: structure, syntactic and statistical pattern Recognition, Springer-Verlag, pp. 806-814, (2004).
24. Kotsiantis, S., Kanellopoulos, D., Pintelas, P. Handling Imbalance Datasets: A review, International Transaction on computer Science and Engineering, 30(1), 25-36, (2006).
25. Napierala, K., Stefanowski, J., Wilk, S. Learning from imbalanced data in presence of noise and Borderline examples: In preceding of the 7th international conference on Rough sets and current Trends in computing, Lecture notes on artificial intelligence vol. 6086, 158-167, (2010).