Generación de Instancias Sintéticas para Clases Desbalanceadas

por

Atlántida Irene Sánchez Vivar

Tesis sometida como requisito parcial para obtener el grado de

MAESTRA EN CIENCIAS EN LA ESPECIALIDAD DE CIENCIAS COMPUTACIONALES

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica

Octubre 2008

Tonantzintla, Puebla

Supervisada por:

Dr. Jesús González Bernal, INAOE
Dr. Eduardo Morales Manzanares, INAOE

Dedicatoria

A mis padres

A mis hermanos y hermanas

A Fernando, el amor de mi vida

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) y al Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) por el soporte económico que me fue otorgado y por los apoyos que recibí durante mis estudios de maestría.

A mi padre, por quien aprendí que para conseguir lo que quieres debes lograrlo por ti mismo. Y a mi madre, que sin importar las circunstancias siempre ha estado conmigo y que con su fortaleza me ha sabido mantener de pie.

A mis hermanos, porque teniéndolos como ejemplo resultan más claras las metas a alcanzar.

A mis sobrinos, que me recuerdan cuando fui como ellos e imaginaba parte de lo que ahora soy. A mis amigos, por brindarme su cariño aún cuando nuestras vidas tomen distintos rumbos.

A mis profesores del INAOE, a mis asesores y mis sinodales. Cada uno de ellos ha contribuido en gran medida con mi formación profesional.

A todo el personal del INAOE.

A Fernando, que me ha entregado su amor. Porque me ayudó a superar los momentos más difíciles y en los mejores estuvo a mi lado para compartir la alegría.

Resumen

Una de las principales dificultades que se presentan en una tarea de clasificación en aprendizaje computacional es el problema de clases desbalanceadas. Este problema se refiere a que, en algunos conjuntos de datos, algunas clases tienen muchos más ejemplos que otras, provocando que el clasificador tienda a aprender más de ellas e ignorar las pequeñas, cuya correcta clasificación generalmente es la de mayor interés. Para fines de esta tesis, se asumen problemas de sólo dos clases: minoritaria y mayoritaria.

Para solucionar este problema se han propuesto varias técnicas, desde las que modifican algoritmos existentes y las que crean nuevos algoritmos hasta las que cambian la distribución de los datos con re-muestreo, todas ellas con la finalidad de favorecer la clasificación de la clase minoritaria. Esta tesis está enfocada en los métodos de re-muestreo, específicamente en el sobre-muestreo de instancias, una técnica que cambia la distribución de datos agregando más instancias de la clase minoritaria y que ha obtenido resultados satisfactorios.

En este trabajo se proponen dos métodos nuevos de sobre-muestreo de instancias, GIS-G y GIS-GF. Ambos métodos parten de la idea de crear grupos de la clase minoritaria y generar las instancias sintéticas dentro de cada grupo, y no de manera global como lo hacen los métodos tradicionales. Además, propone una forma diferente de asignar valores nominales a las nuevas instancias, mientras que los métodos

tradicionales únicamente se enfocan en la asignación de valores numéricos. El primer método, llamado GIS-G, genera nuevos ejemplos interpolando los valores numéricos de pares de instancias dentro de un grupo. El segundo método, llamado GIS-GF, genera los valores de los atributos numéricos de la nueva instancia con sólo una instancia como semilla, haciendo uso de la desviación estándar de los valores dentro del grupo.

Para probar los métodos propuestos se utilizaron veinte bases de datos sintéticas y veintitrés tomadas de dominios reales, se aplicaron los cuatro métodos de sobre-muestreo principales (ROS, SMOTE, Borderline-SMOTE1 y Borderline-SMOTE2) además de los dos métodos propuestos en esta tesis, se utilizó validación cruzada de diez capas sobre cada base de datos, se probaron seis clasificadores diferentes (*AdaBoost M1*, *Naive Bayes*, *K-NN*, *C4.5*, *PART* y *Backpropagation*) y el proceso completo se repitió diez veces para finalmente obtener los promedios de los resultados.

Se mostró, mediante el Análisis ANOVA y pruebas *T*, que los resultados obtenidos por los métodos propuestos presentan en promedio mejores resultados sobre las bases de datos utilizadas, con respecto a los obtenidos por los demás métodos. Estos resultados son estadísticamente significativos.

Palabras clave

Aprendizaje Computacional, Clasificación Supervisada, Clases Desbalanceadas, Sobre-Muestreo.

Abstract

One of the main difficulties that are present in a classification task for machine learning is the problem of unbalanced classes. This problem is that, in some data sets, some classes have many more elements than others, causing the classifier to learn more from them, ignoring the small ones, which right classification is generally of bigger interest. With respect to this thesis, only problems of two classes are assumed (minority and majority).

To solve this problem, many techniques have been proposed, from those methods that that modify existent algorithms and those that create new algorithms, to those that change the distribution of data with resampling, all of them with the objective of favoring the classification of the minority class. This thesis focuses on the resampling methods, specifically on the oversampling of instances, a technique that changes the distribution of data adding more instances of the minority class, that has obtained satisfactory results.

In this work, two new oversampling methods of instances are proposed: GIS-G and GIS-GF. Both methods start from the idea of creating groups of the minority class and generating the new synthetic instances, while the traditional methods are focused only on the numeric values assignment. The first method, named GIS-G, generates new examples by interpolating the numerical values of pairs of instances inside a group. The second method, GIS-GF, generates the values of the numerical attributes of the

instance with just one instance as seed, making use of the standard deviation of the values inside of the group.

To test the proposed methods, twenty databases of synthetic data, and twenty-three databases taken from real domains were used. The four main oversampling methods (ROS, SMOTE, Borderline-SMOTE1 y Borderline-SMOTE2), apart from the methods proposed in this thesis, were applied. Ten-fold cross validation over each database were used. Six different classifiers (*AdaBoost M1*, *Naive Bayes*, *K-NN*, *C4.5*, *PART*, and *Backpropagation*) were tested, and the full process was repeated ten times, to finally obtain the averages of the results.

It was shown, through the ANOVA Analysis and through and T tests, that the obtained results from the proposed methods present, on average, better results over the used databases, with respect to those results obtained by the other methods. These results are estatistically significant.

Key words

Machine Learning, Supervised Classification, Imbalanced Classes, Over-Sampling.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En este capítulo se plantea el problema a tratar, se mencionan los métodos que se han utilizado anteriormente para solucionarlo y se presentan los nuevos métodos propuestos. Además, se listan los clasificadores utilizados para la fase experimental así como las medidas de desempeño utilizadas en la evaluación de los métodos propuestos. Finalmente, se presenta la organización de los siguientes capítulos.

1.1. Planteamiento del problema

El almacenamiento de datos se ha vuelto cotidiano en las actividades de muchos seres humanos. Computadoras con capacidad de memoria cada vez mayor nos ha permitido guardar datos que en otro tiempo habríamos desechado. Actualmente los medios electrónicos permiten registrar las decisiones de la gente: nuestras compras en el supermercado, nuestros hábitos financieros, nuestras preferencias en libros, entre otras cosas. La Internet nos permite el acceso a una inmensa cantidad de información, al mismo tiempo que puede registrar cada elección que hacemos. Todo esto forma parte

importante de la industria y el comercio. Sin embargo, a medida que crece la cantidad de información decrece la capacidad de las personas para entenderla. Información potencialmente útil puede no ser descubierta, por lo que se pierden las ventajas que podrían obtenerse de ella. Hay un gran abismo entre generar datos y entenderlos. Es por ello que surge el concepto conocido como minería de datos (*Data Mining*), que se define como el proceso de descubrir patrones en los datos. Este proceso debe ser automático o (con mayor frecuencia) semi automático. Los patrones descubiertos deben ser representativos de los datos y permitir obtener alguna ventaja, usualmente una ventaja económica.

Las líneas de desarrollo de la minería de datos tiene su origen en tres conceptos fundamentales: la estadística, la inteligencia artificial y el aprendizaje computacional, éste último se puede describir como la unión de las dos primeras. El aprendizaje computacional (*Machine Learning*), que se encarga de estudiar y modelar computacionalmente los procesos de aprendizaje en sus diversas manifestaciones, busca construir programas que mejoren automáticamente con la experiencia. Una de las tareas que se pueden hacer con aprendizaje computacional es la clasificación, donde los datos son objetos caracterizados por atributos que pertenecen a diferentes clases. La meta de la clasificación es inducir un modelo para poder predecir una clase dados los valores de los atributos. Un escenario típico de esta tarea asume la existencia de un conjunto de entrenamiento del cual se induce un clasificador cuyo desempeño se calcula en base a un conjunto de prueba.

Existen varios algoritmos de aprendizaje computacional que han probado ser muy efectivos. Sin embargo, al ser un área en proceso de maduración, se le han presentado retos que no habían sido considerados inicialmente. Uno de esos retos es el problema de clases desbalanceadas, esto es, la distribución de instancias de las diferentes clases es diferente. Este problema de clases desbalanceadas aparece en muchas aplicaciones reales como: detección de llamadas telefónicas con propósito de fraudes

[11], comunicaciones poco confiables para clientes [10], diagnóstico de ciertas enfermedades como la enfermedad tiroidea [2], clasificación de imágenes [24], detección de contracciones intestinales en videos de endoscopías [29], identificación de autoría [25], entre otras. En la mayoría de los casos, la clase de mayor interés es la que tiene menos instancias y esto ocasiona un desempeño pobre en la clasificación de las instancias que pertenecen a esa clase.

El número de clases de un conjunto de datos puede variar, pero para fines de este trabajo se utilizarán sólo dos clases. En caso de haber más de dos clases una se tomará como clase minoritaria o positiva y las demás se fusionarán como una sola clase mayoritaria o negativa.

1.2. Antecedentes

Para resolver el problema de clases desbalanceadas se han propuesto algunos métodos que se pueden dividir en dos grupos principales:

- A nivel algorítmico: modifican algoritmos existentes, asignan diferentes costos a la clasificación o crean nuevos algoritmos. Todo ello con el fin de favorecer la clase minoritaria.
- A nivel de datos: cambian la distribución de los datos para que el clasificador mejore la predicción sobre la clase minoritaria. Se puede hacer con sobre-muestreo (ampliar el conjunto de entrenamiento de la clase minoritaria) o con sub-muestreo (reducir el conjunto de entrenamiento de la clase mayoritaria).

Para el caso de los métodos a nivel algorítmico, una posibilidad es asignar costos a la clasificación de tal forma que se asigne un mayor costo a la mala clasificación de una instancia minoritaria que a la de una mayoritaria. Pero, aunque hay evidencias de

que esta técnica de modificación de costos es efectiva [14], las técnicas propuestas en general dependen de conocimiento del costo asociado a una mala clasificación que no siempre está disponible.

Para el caso de los métodos a nivel de datos puede realizarse un sub-muestreo de la clase mayoritaria con el fin de remover instancias y con ello balancear la distribución de las clases. Una opción es el sub-muestro aleatorio, pero presenta la desventaja de que puede remover instancias importantes para la clasificación. Otro método que puede utilizarse es Tomek-Link [27] que pretende identificar ruido en los datos y así eliminarlo, pero en casos donde la cantidad de ruido sea poca, el problema de desbalance prevalecerá. También existe la técnica CNN [6] que se utiliza para encontrar un subconjunto consistente de instancias y deshecha las demás. Otra opción es la selección de un solo lado (*One-sided selection*) [15], es una versión de CNN y consiste en seleccionar primero un sub conjunto de los ejemplos mayoritarios (utilizando CNN) y conservar todos aquéllos que pertenezcan a la clase minoritaria. Por último, NCL [16] elimina toda instancia cuya etiqueta de clase difiera de la clase de a lo menos dos de sus tres vecinos más cercanos. Sin embargo, CNN, la selección de un solo lado y NCL no solucionan el problema cuando la clase minoritaria es muy pequeña, ya que aunque logran remover ejemplos de la clase mayoritaria aún queda por resolver la falta de ejemplos de la clase minoritaria.

Continuando con los métodos a nivel de datos, el sobre-muestreo permite hacer más grande la clase minoritaria y así permitir al clasificador aprender mejor sobre ella. Aunque lo más fácil y directo es realizar un sobre-muestreo aleatorio, éste puede producir sobre-ajuste. A partir de este punto surgen los métodos “inteligentes” que buscan generar instancias sintéticas a partir de las instancias orginales pero no repitiendo valores. Dentro de estos métodos inteligentes se encuentra SMOTE (por sus siglas en inglés *Synthetic Minority Over Sampling Technique*) [4], que crea nuevas instancias interpolando entre pares de instancias de la clase minoritaria cercanas entre sí. Tam-

bién existen *Borderline-SMOTE1* y *Borderline-SMOTE2* [13], que están basados en SMOTE, con la diferencia de que sólo interpolan entre instancias que se consideren en el borde de la clase. Borderline-SMOTE1 interpola con ejemplos de la clase minoritaria mientras que Borderline-SMOTE2 lo hace con instancias de los bordes de ambas clases. A pesar de que estos métodos han mostrado su eficacia, aún hay aspectos que no consideran y que los lleva a un desempeño de clasificación bajo en algunas situaciones: en el caso de SMOTE, al interpolar entre instancias pueden existir instancias de la clase mayoritaria entre ellas, provocando la generación de modelos incorrectos; los métodos Borderline-SMOTE1 y Borderline-SMOTE2 no siempre son capaces de encontrar los bordes, por lo que no se puede llevar a cabo el sobre-muestreo y de esa manera no se mejora la clasificación de la clase minoritaria. Además, estos métodos inteligentes sólo trabajan con atributos numéricos (los atributos nominales se tratan de una forma simple) y no tienen una forma clara de definir cuántos vecinos más cercanos utilizar.

Aunque se han implementado diversos métodos para tratar con el problema de clases desbalanceadas, los resultados obtenidos aún no pueden generalizarse para decidir qué método resulta más conveniente. Esto se debe a que es un problema relativo a la complejidad del concepto (lo que es aprendido), tamaño del conjunto de datos y nivel de desbalance [14]. Además, como ya se mencionó, los métodos propuestos presentan algunos inconvenientes. En este trabajo se opta por trabajar con sobre-muestreo, tomando en cuenta las debilidades de los metodos anteriores y las siguientes consideraciones:

- Sobre-muestreo es la clase de método que ataca más directamente el problema de la falta de instancias de la clase minoritaria [1].
- Sobre-muestreo produce el mismo efecto que mover los umbrales de decisión o ajustar costos [17], pero es independiente del algoritmo de clasificación.
- SMOTE y Borderline-SMOTE han mostrado buenos resultados pero necesitan

considerar condiciones extras para un mejor desempeño.

- Se requiere de una técnica para tratar los valores de los atributos nominales.

1.3. Solución propuesta

Dadas las condiciones de la sección anterior, los principales retos computacionales a enfrentar son:

- El proceso de generación de instancias sintéticas debe tener como resultado instancias válidas, en el sentido de que éstas deben ser generadas en áreas que estén mayormente influenciadas por la clase minoritaria. Se debe evitar crear instancias que provoquen un mayor grado de traslape (*overlap*) entre clases.
- La generación de valores para atributos nominales debe ser pensada de tal manera que no haya sesgo en ellos como sucede en el caso de SMOTE, Borderline-SMOTE1 y Borderline-SMOTE2. Este sesgo se debe a que la asignación de un valor nominal se lleva a cabo considerando el voto mayoritario de los vecinos más cercanos (el valor que más se repite entre los vecinos es el que se asigna a la nueva instancia). Entonces, a medida que crezca el porcentaje de sobre-muestreo el sobre-ajuste en los valores nominales también aumentará, porque el voto mayoritario de los vecinos más cercanos siempre dará el mismo resultado.
- La definición de cuántos vecinos más cercanos considerar para generar las instancias sintéticas no debe ser un parámetro arbitrario, debe determinarse de acuerdo a las características de los datos en consideración.

En esta tesis se proponen dos algoritmos de sobre-muestreo inteligente como herramientas para resolver el problema de clases desbalanceadas considerando los retos computacionales que se presentaron en el párrafo anterior. Estos métodos, llamado

GIS-G y GIS-GF, crean grupos de instancias de la clase minoritaria y generan las instancias sintéticas dentro de cada grupo, así las instancias sintéticas tendrán una validez mayor que si se generaran de forma global. Para generar los valores de los atributos nominales lo hace de acuerdo a la proporción de valores de los atributos que hay en el grupo, con esto se evita la repetición de valores que se daría usando voto mayoritario de los vecinos más cercanos. Para crear un valor numérico GIS-G lo hace por interpolación (igual que SMOTE), pero GIS-GF lo hace utilizando la desviación estándar y desviación media de los datos y tomando como semilla una sola instancia. Para el caso de la definición de cuántos vecinos más cercanos considerar, GIS-G y GIS-GF toman en cuenta el número de instancias que hay en un grupo y no un parámetro k ya establecido. Aunque los métodos pueden aplicarse a problemas con más de dos clases y se puede sobre-muestrear más de una clase, para fines experimentales se consideran problemas con dos clases, se toma una como minoritaria y es llamada la clase positiva, las demás se unen como una sola clase mayoritaria y es llamada la clase negativa.

Para la fase de evaluación se utilizaron veinte bases de datos generadas artificialmente y veintitrés tomadas de dominios reales. Se hizo validación cruzada en diez capas aplicada sobre seis diferentes clasificadores (ensamble de clasificadores, probabilista, vecinos más cercanos, árbol de decisión, reglas y redes neuronales). Debido a que la métrica de desempeño exactitud (*accuracy*) resulta injusta en dominios de clases desbalanceadas, se utilizaron las medidas recuerdo (*True Positive Rate* o *Recall*), precisión (*precision*), Medida-F (*F-measure*) y AUC (*Area Under ROC Curve*). Para probar la eficacia de los métodos propuestos en esta tesis sobre los métodos de sobre-muestreo propuestos en el estado del arte se llevaron a cabo pruebas de significancia estadística así como análisis de varianza, los cuales mostraron que GIS-G y GIS-GF son mejores que los otros métodos de sobre-muestreo. Esta mejora se ve de más claramente cuando la comparación de resultados se realiza método a método.

1.4. Organización de la tesis

El siguiente capítulo presenta los conceptos básicos involucrados en este trabajo de investigación. El capítulo 3 reporta los métodos para tratar el problema de clases desbalanceadas. El capítulo 4 describe los dos nuevos métodos propuestos para la solución al problema de clases desbalanceadas, trabajando sobre los puntos débiles de los métodos de sobre-muestreo conocidos. En el capítulo 5 se detallan las bases de datos utilizadas para los experimentos realizados, se establece una forma de evaluar los métodos desarrollados y compararlos con los ya existentes y se muestran, analizan y discuten los resultados obtenidos. Las conclusiones que se derivan de este trabajo y el trabajo a desarrollar que se desprende de ella se especifican en el capítulo 6.

Capítulo 2

Marco Teórico

En este capítulo se definen los conceptos básicos involucrados en este trabajo de investigación. En particular, se explica el problema de clases desbalanceadas, se presentan las funciones de distancia utilizadas para determinar la vecindad entre instancias, se describe el proceso de normalización utilizado para unificar el rango de valores que toman los atributos numéricos y por último se describe el Análisis de Varianza así como las pruebas de significancia estadística.

2.1. Conceptos básicos

2.1.1. Minería de Datos y Aprendizaje Computacional

La minería de datos (*Data mining*) se encarga de buscar patrones en los datos. Este tema no es reciente, ha existido desde el comienzo de la humanidad. Los cazadores conocían el comportamiento de sus presas, los agricultores buscaron patrones en el cultivo de la tierra, los políticos han buscado patrones en los votantes, gran parte de

nuestra vida está basada en los patrones que hemos aprendido. El trabajo de un científico en esta área es darle sentido a los datos, descubrir los patrones de cómo trabaja el mundo físico y convertirlos en teorías. Éstas podrán entonces ser utilizadas para predecir qué sucederá en situaciones nuevas.

Por ejemplo, supongamos que se tiene una base de datos de las compras realizadas en un centro comercial. Al encontrar los patrones en estas compras, pueden planearse de manera más eficiente los servicios que se proporcionen a los clientes, los productos que se ofertarán y la ubicación de ellos. Si los datos se analizan inteligentemente son un recurso muy valioso. En la minería de datos, los datos se almacenan electrónicamente y la búsqueda así como el análisis son automáticos. Los economistas, estadísticos, ingenieros en comunicaciones, entre otros, han trabajado por mucho tiempo sobre la idea de cómo descubrir patrones en los datos. En los últimos años ha habido un gran crecimiento de las técnicas utilizadas para encontrar patrones en los datos.

Descripción de patrones

Para poder encontrar patrones, es necesario que los datos se presenten en una forma determinada. Vea la tabla 2.1 donde se dan las condiciones bajo las cuales un oculista podría prescribir o no lentes de contacto. Cada línea de la tabla es uno de los *ejemplos*. Asimismo, una representación basada en reglas sería como sigue:

if Tasa de producción de lágrimas = reducida **then**
  recomendación = ninguna
else if Etapa = temprana and astigmatismo = no **then**
  recomendación = soft
end if

Etapa	**Prescripción de gafas**	**Astigmatismo**	**Tasa de producción de lágrimas**	**Recomendación de lentes**
temprana	miopía	no	reducida	ninguna
temprana	miopía	no	normal	suave
temprana	miopía	yes	reducida	ninguna
temprana	miopía	yes	normal	fuerte
temprana	hipermetropía	no	reducida	ninguna
temprana	hipermetropía	no	normal	suave
temprana	hipermetropía	yes	reducida	ninguna
temprana	hipermetropía	yes	normal	fuerte
pre-presbicia	miopía	no	reducida	ninguna
pre-presbicia	miopía	no	normal	suave
pre-presbicia	miopía	yes	reducida	ninguna
pre-presbicia	miopía	yes	normal	fuerte
pre-presbicia	hipermetropía	no	reducida	ninguna
pre-presbicia	hipermetropía	no	normal	suave
pre-presbicia	hipermetropía	yes	reducida	ninguna
pre-presbicia	hipermetropía	yes	normal	ninguna
presbicia	miopía	no	reducida	ninguna
presbicia	miopía	no	normal	ninguna
presbicia	miopía	yes	reducida	ninguna
presbicia	miopía	yes	normal	fuerte
presbicia	hipermetropía	no	reducida	ninguna
presbicia	hipermetropía	no	normal	suave
presbicia	hipermetropía	yes	reducida	ninguna
presbicia	hipermetropía	yes	normal	ninguna

Tabla 2.1: La base de datos *Lentes de contacto*

Las descripciones no necesariamente deben estar en esta forma. Los árboles de decisión, que especifican las secuencias de decisiones que necesitan ser hechas y la recomendación resultante, son otros medios populares de expresión.

Este ejemplo es simple. Hay 24 renglones, representando tres valores posibles para *Etapa* y dos valores para *Prescricpión de gafas*, *Astigmatismo*, y *Tasa de producción de lágrimas* ($3 \times 2 \times 2 \times 2 = 24$ que es el total de renglones).

Primero, en este caso todas las combinaciones de valores posibles están en la tabla. Sin embargo, en la mayoría de las situaciones de aprendizaje el conjunto de ejemplos de entrada tiene datos faltantes y muy probablemente ruido, por lo que parte del trabajo es generalizar a ejemplos nuevos. Se pueden omitir algunos de los renglones de la tabla 2.1 para la que *Tasa de producción de lágrimas* sea *reducida* y aún tener la regla

if Tasa de producción de lágrimas = reducida **then**
  recomendación = ninguna
end if

con la que podría predecirse la clase de los renglones faltantes en forma correcta. Segundo, los valores se especifican para todas las características en todos los ejemplos. Muchos conjuntos de datos del mundo real contienen ejemplos en los que los valores de algunas características, por una u otra razón, son desconocidos. Tercero, las reglas precedentes clasifican los ejemplos correctamente, en tanto que frecuentemente, por los errores de ruido en los datos, la mala clasificación ocurre incluso con datos utilizados en el proceso de entrenamiento del clasificador.

Aprendizaje computacional

Dentro de la minería de datos se utilizan algoritmos de aprendizaje computacional (*Machine Learning*). Ahora que se tiene idea acerca de las entradas y salidas, veamos sobre estos algoritmos. Para empezar, ¿qué es el aprendizaje? ¿qué es el aprendizaje computacional?. Éstas son preguntas filosóficas, por lo que se tratará de un modo práctico. Una definición operacional de aprendizaje se puede formular de la misma manera:

Algo aprende cuando es capaz de cambiar su comportamiento de tal forma que su desempeño sea mejor en el futuro.

Ésta es una definición objetiva. Se puede probar el aprendizaje observando el comportamiento y comparándolo con el comportamiento pasado. Existen diversas tareas que se pueden llevar a cabo con sistemas de aprendizaje computacional y se listan a continuación:

- Descripción: normalmente se usa como análisis preliminar de los datos y busca derivar descripciones concisas de características de los datos.
- Predicción: se divide en Clasificación y Estimación.
    - Clasificación: los datos son objetos caracterizados por atributos que pertenecen a diferentes clases (etiquetas discretas). La meta es inducir un modelo para poder predecir una clase dados los valores de los atributos.
    - Estimación o Regresión: las clases son continuas. La meta es inducir un modelo para poder predecir el valor de la clase dados los valores de los atributos.
- Segmentación: separación de los datos en sub-grupos o clases interesantes. Éstas pueden ser exhaustivas y mutuamente exclusivas o jerárquicas y con traslapes.

- Análisis de dependencias: el valor de un elemento puede usarse para predecir el valor de otro. La dependencia puede ser probabilista, puede definir una red de dependencias o puede ser funcional (leyes físicas).
- Detección de desviaciones, casos extremos o anomalías: detectar los cambios más significativos en los datos con respecto a valores pasados o normales. Sirve para filtrar grandes volúmenes de datos que son menos probables de ser interesantes.
- Aprendizaje por refuerzo: consiste en aprender a decidir, ante una situación determinada y a partir de la experiencia, que acción es la más adecuada para lograr un objetivo.
- Optimización y búsqueda: existen una gran cantidad de algoritmos de búsqueda tanto determinística como aleatoria, individual como poblacional, local como global, que se utilizan principalmente para resolver algún problema de optimización. Aquí podemos incluir a los algoritmos genéticos, recocido simulado, ant-colony, técnicas de búsqueda local, etc.

Esta tesis está enfocada a la tarea de predicción, específicamente en clasificación.

2.1.2. Conceptos, Instancias y Atributos

Para la operación de los métodos de aprendizaje computacional, la entrada toma la forma de *conceptos*, *instancias*, y *atributos*.

Concepto

Llamamos *concepto* a lo que es aprendido y *descripción del concepto* a la salida producida por el esquema de aprendizaje. La idea de un concepto, como la sola idea

de aprendizaje en primer lugar, es difícil de definir precisamente, y se gastará tiempo filosofando sólo acerca de lo que es o no es. En un sentido, lo que estamos tratando de encontrar, el resultado de un proceso de aprendizaje, es una descripción del concepto que es *inteligible* en la medida que puede ser entendido, discutido, y disputado, y *operacional* en la medida que puede ser aplicado a instancias actuales. Por ejemplo, dado el problema de la tabla 2.1, el objetivo es aprender cómo decidir acerca de la recomendación del uso de lentes a un paciente nuevo.

Para representar un concepto, como lo hace Mitchel, se usa un lenguaje de representación basado en marcos. Por ejemplo, el concepto *coche* se representa:

origen: x1

marca: x2

color: x3

década: x4

tipo: x5

x1 pertenece a Japón, EEUU, UK, Italia,...

x2 pertenece a Honda, Toyota, Chrysler, Fiat,...

x3 pertenece a azul, blanco, amarillo, verde,...

x4 pertenece a 1950,1960,1970,1980,1990,2000,...

x5 pertenece a económico, lujo, deportivo,...

La descripción de conceptos se puede poner en términos de ranuras y valores. Por ejemplo, el concepto coche económico japonés

origen = Japón

marca = x2

color = x3

década = x4

tipo = económico

Instancia

La información que se le da al clasificador para la fase de aprendizaje toma la forma de un conjunto de *instancias*, en la siguiente sección se explica lo que es un clasificador. Aunque anteriormente se han estado llamando *ejemplos*, en adelante se utilizará el término más específico *instancias* para referirse a la entrada. Cada instancia es un ejemplo individual e independiente del concepto a ser aprendido. Un ejemplo o instancia del concepto *coche* en particular es:

origen = Japón

marca = Honda

color = azul

década = 1970

tipo = económico

Atributo

Cada instancia se caracteriza por los valores de un conjunto predeterminado de atributos que miden diferentes aspectos de la instancia. Hay varios tipos de atributos, éstos pueden ser divididos en dos grupos: *numéricos* y *nominales* (o *categóricos*). Los atributos numéricos, algunas veces llamados atributos *continuos*, miden números (enteros o reales). Los atributos nominales toman valores en un conjunto específico y finito de posibilidades. Las cantidades nominales tienen valores que son distintos símbolos, los valores por sí solos sirven sólo como etiquetas o nombres (de ahí el término *nominal*, que viene de la palabra en latín para nombre).

2.2. Clasificadores

Los métodos de clasificación son llamados *clasificadores* y existe una gran variedad de ellos [32]. Éstos pueden agruparse en cinco tipos diferentes: bayesianos (probabilistas), de generación de árboles de decisión, de generación de reglas, de funciones y los llamados flojos (métodos simples como el de vecinos más cercanos). Para este trabajo se seleccionó un clasificador de cada tipo más un ensamble de clasificadores, éstos se describen a continuación:

- Clasificador Bayesiano Simple (*Naive Bayes*): Es un clasificador probabilista simple basado en la aplicación del Teorema de Bayes que asume independencia entre los atributos dada la clase.
- C4.5: Genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente. El árbol se construye mediante la estrategia (*hill-climbing*).
- PART: Con C4.5 genera listas de decisión mediante reglas de la forma *if-then*.

- Retropropagación (*Backpropagation*): Construye una red neuronal donde se interconectan varias unidades de procesamiento en capas, las neuronas de cada capa no se conectan entre sí. Sin embargo, cada neurona de una capa proporciona una entrada a cada una de las neuronas de la siguiente capa, esto es, cada neurona transmitirá su señal de salida a cada neurona de la capa siguiente.

- K-vecinos más cercanos (K-NN): Con este método se clasifica una instancia mediante el voto mayoritario de sus vecinos.

- *Adaptive Boosting* (*AdaBoost M1*) [12]: Construye clasificadores mediante la asignación de pesos a los ejemplos de forma adaptativa. En cada iteración de *boosting* se construye un clasificador que intenta compensar los errores cometidos previamente por otros clasificadores. Para lograr que cada nuevo clasificador mejore los resultados en regiones donde fallan los anteriores se utiliza un conjunto de datos ponderado cuyos pesos son actualizados tras cada iteración, se incrementan los pesos de los ejemplos mal clasificados por el último clasificador y se reducen los pesos de los bien clasificados.

2.3. Problema de clases desbalanceadas

El objetivo del aprendizaje computacional es desarrollar métodos computacionales que implementan mecanismos capaces de inducir conocimiento a partir de un conjunto de datos, ya que los datos por sí solos muchas veces no producen un beneficio directo. Sin embargo, estos métodos aplicados a problemas reales han traído una serie de retos que no habían sido considerados cuando se crearon. Uno de tales retos es el problema de clases desbalanceadas, y aunque ha cobrado mucho interés recientemente en la comunidad científica aún hay mucho trabajo por hacer.

En un problema de clasificación, el problema de desbalance ocurre típicamente cuando hay muchas más instancias de una clase que de otra [5]. En tales casos, la tendencia en los clasificadores estándares es aprender mejor de la clase que tiene más instancias e ignorar la que menos tiene, ya que al contar con pocos datos para una clase es muy difícil para un clasificador encontrar patrones que los cubra con el suficiente soporte estadístico como para considerarlos. Esto implica un desempeño bastante pobre en la clasificación de la clase minoritaria, sobre todo porque esta clase generalmente es la de mayor interés. Al haber menos instancias con que entrenar al clasificador será difícil encontrar suficientes patrones que definan esa clase. De ahí que surge la necesidad de diseñar e implementar métodos computacionales que traten ese problema.

Como se verá en el capítulo 3, se han propuestos varias opciones para solucionar este problema. Estos métodos se pueden dividir en dos grupos: los que cambian la distribución de datos con sub-muestreo o sobre-muestreo y los que crean nuevos algoritmos o modifican los ya existentes. Ambos con la finalidad de mejorar la clasificación de la clase minoritaria. Esta tesis se enfoca a los métodos de sobre-muestreo.

2.4. Funciones de distancia

Casi todos los métodos de sobre-muestreo del Capítulo 3 (excepto ROS) requieren conocer los vecinos más cercanos de una instancia dada, ya sea de una u otra clase. Pero para poder estimar esa cercanía es necesario conocer la distancia que existe entre dichas instancias. Las funciones de distancia se utilizan para determinar qué tan cercano está un miembro de un conjunto de instancias a otra instancia dada. Una vez que se tiene esa estimación pueden hacerse varias cosas como: predecir la clase de la instancia de acuerdo a sus vecinos más cercanos, generar nuevos ejemplos entre esa instancia y sus vecinos, generar grupos de instancias de una misma clase, entre otras.

El problema es definir la función de distancia, y esto se debe hacer tanto para instancias con atributos numéricos, nominales o con ambos. El caso de los atributos nominales es un caso especial en el sentido de que el uso del término *distancia* no es adecuado, éste se refiere a una medida entre valores numéricos. Lo adecuado sería decir *similaridad* o *dissimilaridad* entre los valores nominales de un atributo. Pero, dado que existe el caso de que haya instancias con atributos de ambos tipos, por cuestiones prácticas utilizaremos el término *distancia* tanto para atributos numéricos como nominales. Un resumen de las funciones de distancia más utilizadas para atributos numéricos se presenta en la tabla 2.2. Aunque hay varias opciones, la mayoría de los métodos de clasificación, incluídos los de este trabajo, utilizan la distancia euclidiana.

Para el caso de atributos numéricos el cálculo de distancia puede resultar trivial, ya que pueden establecerse funciones lineales para ellos. Sin embargo, existe el problema de estimación de distancia para atributos nominales, con los cuales el uso de una medida de distancia lineal tiene poco sentido por lo que es necesario utilizar otro tipo de medidas. Además, se debe considerar que en muchos problemas aparecen ambos tipos de atributos.

Una opción para el manejo de aplicaciones con tipos de atributos numéricos y nominales es utilizar una función de distancia heterogénea que utilice diferentes funciones de distancia sobre diferentes tipos de atributos [31]. Esto es, una función para atributos numéricos y otra para atributos nominales. Estas funciones se describen a continuación.

2.4.1. HEOM (Métrica Heterogénea Euclidiana-Traslape)

En esta métrica [31] se contemplan dos casos: cuando el atributo es de tipo numérico y cuando es de tipo nominal. Éstos son tratados de diferente manera para estimar la distancia o disimilaridad entre dos ejemplos. Un enfoque que ha sido utili-

Minkowsky: $D(x,y) = (\sum_{i=1}^{m} \lvert x_i - y_i \rvert^r)^{1/r}$	**Euclidiana:** $D(x,y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$
Manhattan: City-block: $D(x,y) = \sum_{i=1}^{m} \lvert x_i - y_i \rvert$	**Camberra:** $D(x,y) = \sum_{i=1}^{m} \left\lvert \frac{x_i - y_i}{x_i + y_i} \right\rvert$
Chebychev: $D(x,y) = \operatorname*{máx}_{i=1}^{n} \lvert x_i - y_i \rvert$	**Mahalanobis:** $D(x,y) = [detV]^{1/m}(x-y)^T V^{-1}(x-y)$ donde V es la matriz de covarianza de los atributos y det es el determinante de la matriz.
Cuadrática: $D(x,y) = (x-y)^T Q(x-y) = \sum_{j=1}^{m}(\sum_{i=1}^{m}(x_i - y_i)q_{j,i})(x_j - y_j)$ donde Q es una matriz de pesos definida minoritaria, específica del problema.	
Correlación $D(x,y) = \frac{\sum_{i=1}^{m}(x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x}_i)^2 \sum_{i=1}^{m}(y_i - \bar{y}_i)^2}}$ donde $\bar{x}_i$ y $\bar{y}_i$ es el valor medio de los respectivos atributos.	
Chi-cuadrada: $D(x,y) = \sum_{i=1}^{m} \frac{1}{sum_i}(\frac{x_i}{size_x} - \frac{y_i}{size_y})^2$ donde sum_i es la suma de todos los valores del atributo i que aparecen en la matriz de entrenamiento, y $size_x$ es la suma de todos los valores del vector x.	
Correlación de ranking de Kendall: $D(x,y) = 1 - \frac{2}{n(n-1)} \sum_{i=1}^{m} \sum_{j=1}^{i-1} sign(x_i - x_j) sign(y_i - y_j)$ donde $sign$ es una función que transforma los negativos y positivos a {-1, 1}	

Tabla 2.2: Funciones de distancia más utilizadas para valores numéricos. Las instancias *x* y *y* son descritas por *m* atributos. x_i y y_i son los valores de i-ésimo atributo de los objetos *x* y *y*, respectivamente.

zado es usar la distancia euclidiana para atributos numéricos y la métrica de traslape para atributos nominales. La distancia total entre dos instancias se obtiene de la sumatoria de las distancias entre cada par de atributos.

- Numéricos: se calcula distancia euclidiana con la fórmula que se muestra de la tabla 2.2.
- Nominales: se calcula de la siguiente manera:

$$traslape(x, y) = \begin{cases} 0 & \text{si x = y} \\ 1 & \text{en otro caso} \end{cases} \tag{2.1}$$

2.4.2. Métrica Heterogénea de Diferencia de Valores

Esta métrica [31] obtiene la distancia entre dos instancias x y y como sigue:

$$HVDM(x, y) = \sqrt{\sum_{a=1}^{m} d_a^2(x_a, y_a)} \tag{2.2}$$

Donde:

- m es el número de atributos.
- La función $d_a(x, y)$ regresa una distancia entre los dos valores x y y para el atributo a y se define como:

$$d_a(x, y) = \begin{cases} vdm(x, y) & \text{si } a \text{ es nominal} \\ Euclidiana(x, y) & \text{si } a \text{ es numérico} \end{cases} \tag{2.3}$$

VDM fue introducida en 1986 por Stanfill y Waltz para proporcionar una función de distancia apropiada para atributos nominales. Una versión simplificada de VDM

define la distancia entre dos valores x y y de un atributo a como:

$$vdm_a(x,y) = \sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^q = \sum_{c=1}^{C} |P_{a,x,c} - P_{a,y,c}|^q \tag{2.4}$$

Donde:

- $N_{a,x}$ es el número de instancias en el conjunto de entrenamiento T que tiene el valor x para el atributo a;
- $N_{a,x,c}$ es el número de instancias en T que tienen el valor x para el atributo a y la clase es c;
- C es el número de clases en el dominio del problema;
- q es una constante, usualmente 1 o 2; y
- $P_{a,x,c}$ es la probabilidad condicional de que la clase sea c dado que el atributo a tiene el valor x, es decir, $P(c|a_x)$. $P_{a,x,c}$ está definida como:

$$P_{a,x,c} = \frac{N_{a,x,c}}{N_{a,x}} \tag{2.5}$$

Donde $N_{a,x}$ es la suma de $N_{a,x,c}$ sobre todas las clases, es decir:

$$N_{a,x} = \sum_{c=1}^{C} N_{a,x,c} \tag{2.6}$$

y la suma de $P_{a,x,c}$ sobre todas las C clases es 1 para valores fijos de a y x.

Utilizando la medida de distancia VDM, dos valores se consideran muy cercanos si tienen clasificaciones similares (correlación similar con la clase de salida). En esta tesis se eligió HVDM para estimar la distancia entre instancias.

2.5. Normalización

Como se mencionó, la distancia entre atributos numéricos se hace generalmente con la función de distancia euclidiana. Sin embargo, una debilidad de esta función es que si uno de los atributos de entrada tiene un rango relativamente amplio, entonces puede dominar a los otros atributos. Por ejemplo, si una aplicación tiene sólo dos atributos, A y B, y A puede tomar valores entre 1 y 1000, y los valores de B pueden ir sólo de 1 a 10, entonces la influencia de los valores de B sobre la función de distancia será frecuentemente dominada por la influencia de los valores de A. Por tanto, las distancias necesitan ser normalizadas dividiendo la distancia de cada atributo entre el rango (máximo - mínimo) de ese atributo, de modo que la distancia de cada atributo esté en el rango de 0 a 1. Los datos necesitan ser normalizados antes de aplicar cualquier método y así no es necesario normalizar el resultado de la función de distancia. La función de normalización que se utiliza es la siguiente:

$$nuevo\_valor_a = (valor_a - valor_{min})/(valor_{max} - valor_{min}) \tag{2.7}$$

Donde:

- $valor_a$ es el valor de atributo que se desea normalizar.
- $valor_{min}$ es el valor mínimo que puede tomar el atributo a.
- $valor_{max}$ es el valor máximo que puede tomar el atributo a.

2.6. Medidas de desempeño

En un problema de clasificación generalmente nos interesa saber el porcentaje de datos clasificados correctamente. Pero el desempeño de un clasificador que fue entrenado con clases desbalanceadas no debe ser medido en términos de la exactitud (*accuracy*, porcentaje de ejemplos de prueba reconocidos correctamente por el clasificador), ya que al ser una medida global de desempeño resulta una medida injusta cuando se trata de clases desbalanceadas. Para explicar lo anterior utilicemos el siguiente ejemplo: Se tiene un dominio con 100 instancias, de las cuales 90 pertenecen a la clase mayoritaria (negativa) y 10 a la clase minoritaria (positiva), y supongamos que un clasificador que al ser entrenado con estos datos predice bien sólo los ejemplos de la clase mayoritaria. Entonces la exactitud sería igual a 0.9, si consideramos que el valor más alto que puede tomar esta medida es 1 entonces se concluiría erróneamente que el desempeño del clasificador es muy bueno, lo cual es falso porque la clase minoritaria (la de mayor interés) ha sido completamente mal identificada aún cuando la mayoritaria ha sido clasificada muy bien. Por lo anterior debe optarse por tomar otras medidas para evaluar el desempeño de los clasificadores cuando se trabaja con clases desbalanceadas.

La tabla 2.3 muestra la matriz de confusión para un problema de dos clases. La primera columna de la tabla es la etiqueta verdadera de la clase de las instancias, y el primer renglón presenta la etiqueta de clase que se predijo. TP (*True Positive*) y TN (*True Negative*) denotan el número de instancias positiva (minoritarias) y negativas (mayoritarias), respectivamente, que se clasificaron correctamente, mientras que FN (*False Negative*) y FP (*False Positive*) denotan el número de instancias positivas y negativas, respectivamente, que fueron mal clasificadas.

	Se predijo positivo	**Se predijo negativo**
Es positivo	TP	FN
Es negativo	FP	TN

Tabla 2.3: Matriz de confusión para un problema de dos clases.

A continuación se listan las medidas de desempeño que se obtienen de esta tabla:

- Exactitud (*Accuracy*)

$$exactitud = (TP + TN)/(TP + FN + FP + TN) \tag{2.8}$$

- Tasa de Falsos Positivos (*False Positive Rate*)

$$FPR = FP/(TN + FP) \tag{2.9}$$

- Tasa de Verdaderos Positivos o Recuerdo (*True Positive Rate* o *Recall*)

$$recuerdo = TP/(TP + FN) \tag{2.10}$$

el recuerdo tiene la ventaja de ser independiente del costo de clasificación y las probabilidades *a priori*.

- Precisión

$$precision = TP/(TP + FP) \tag{2.11}$$

- Medida-F (*F-Measure*)

$$Medida - F = ((1 + \beta^2) * recuerdo * precision)/(\beta^2 * recuerdo + precision) \tag{2.12}$$

Medida-F es alta cuando precisión y recuerdo son altos. β corresponde a la importancia relativa de precisión vs recuerdo, usualmente se le asigna valor 1.

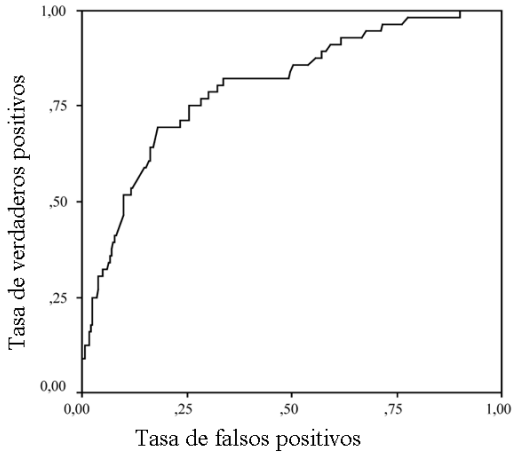

Figura 2.1: Ejemplo de una curva ROC.

- Curvas ROC: Son una de las más populares métricas para evaluar clasificadores, es una medida para determinar qué tan bueno es un algoritmo de clasificación, esto lo hace graficando cierta curva y midiendo el área bajo ésta. Es una gráfica bidimensional donde en el eje y se traza recuerdo y en el eje x se traza la tasa de FPR. FPR denota el porcentaje de ejemplos mal clasificados y el recuerdo es el porcentaje de ejemplos correctamente clasificados. El punto (0, 1) en la curva ROC es el punto ideal para los clasificadores. La curva ROC representa una compensación relativa entre beneficio (recuerdo) y costo (FPR). AUC especifica la probabilidad de que, cuando sacamos aleatoriamente un ejemplo positivo y uno negativo, la función de decisión asigna un valor más alto al positivo que al negativo. Un ejemplo de estas curvas se muestra en la figura 2.1.

- AUC (Área Bajo la Curva ROC) también se puede aplicar a esta evaluación. Es el mejor indicador global de la precisión de una prueba diagnóstica. Hace factible expresar el desempeño de una prueba mediante un número simple.

Para evaluar el desempeño de los métodos propuestos en esta tesis se eligieron las medidas recuerdo, precisión, Medida-F y AUC.

2.7. Evaluación de resultados experimentales

2.7.1. Validación cruzada

Cuando se tiene que probar el desempeño de métodos de aprendizaje computacional sobre un conjunto de datos, es muy importante que los datos utilizados para crear el clasificador (datos de entrenamiento) no sean los mismos para probar el desempeño de éste (datos de prueba). Pero esos conjuntos, de entrenamiento y prueba, deben seleccionarse de tal forma que se garantice que cada clase está representada proporcionalmente en cada uno de ellos sin sesgar los datos. A esto se le llama *estratificación*. Una manera general de disminuir el sesgo causado por elección de las muestras y la aleatoriedad en los métodos de aprendizaje computacional es repetir el proceso completo de entrenamiento y prueba varias veces y con diferentes muestras. La técnica estadística más aceptada para realizar esto es la *validación-cruzada*, en la cual se decide un número fijo de *capas* o particiones de los datos. Por ejemplo, supongamos que se establecen diez particiones, entonces el conjunto de datos se divide en diez partes aproximadamente iguales y cada parte en turno se utiliza para prueba y las restantes para entrenamiento. El procedimiento se repite para cada una de las diez partes y finalmente se obtiene el promedio de los resultados. Esto se conoce como validación cruzada en diez capas (en inglés *ten-fold cross validation*), que en términos prácticos ha llegado a ser un método estándar.

2.7.2. Significancia estadística

En muchos problemas de la ciencia es necesario decidir si se acepta o se rechaza un enunciado acerca de algún parámetro. Al enunciado se le llama *hipótesis* y al procedimiento para tomar decisiones acerca de la hipótesis se le llama *prueba de hipótesis*. Se considera la prueba estadística de hipótesis [20] como la etapa de análisis

de datos de un *experimento comparativo*, en el cual nos interesa comparar la media de unos resultados con la media de otros. En particular, estamos interesados en saber si los resultados de los métodos de sobre-muestreo propuestos en esta tesis son significativamente mejores que los reportados en el estadao del arte. En este caso se consideran las inferencias estadísticas sobre la diferencia de las medias, μ_1 y μ_2, de los resultados de dos métodos de sobre-muestreo, uno comparado contra el otro, donde las varianzas σ_1 y σ_2 son conocidas. Los supuestos que se utilizarán en esta sección se resumen a continuación:

1. $X_{11}, X_{12}, ..., X_{1n}$ es una muestra aleatoria de la población 1.

2. $X_{21}, X_{22}, ..., X_{2n}$ es una muestra aleatoria de la población 2.

3. Las poblaciones de los dos puntos anteriores se representan por X_1 y X_2 y son independientes.

4. Ambas poblaciones son normales, o si no lo son, se cumplen las condiciones del teorema del límite central. Tal teorema indica que, bajo condiciones muy generales, la distribución de la suma de variables aleatorias tiende a una distribución normal cuando la cantidad de variables es muy grande.

Un estimador puntual lógico de $\mu_1 - \mu_2$ es la diferencia de las medias muestrales $\bar{X}_1 - \bar{X}_2$. Suponga que el interés está en probar que la diferencia de las medias μ_1 y μ_2 es igual a un valor especificado Δ_0. Por tanto, la hipótesis nula se enunciará como

$$H_0 : \mu_1 - \mu_2 = \Delta_0 \tag{2.13}$$

y la hiótesis alternativa como

$$H_1 : \mu_1 - \mu_2 \neq \Delta_0 \tag{2.14}$$

Ahora bien, un valor muestral de $\bar{x}_1 - \bar{x}_1$ que difiera considerablemente de Δ_0 es evidencia de que H_1 es verdadera (hay significancia estadística en los resultados de los métodos de sobre-muestreo). A continuación se hace el resumen formal de estos resultados.

Prueba de hipótesis sobre $\mu_1 - \mu_2$, varianzas conocidas

Hipótesis nula:

$$H_0 : \mu_1 - \mu_2 = \Delta_0 \tag{2.15}$$

Estadístico de la prueba:

$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{2.16}$$

Hipótesis alternativa

$$H_1 : \mu_1 - \mu_2 \neq \Delta_0 \tag{2.17}$$

Criterio de rechazo

$$z_0 > z_\alpha \tag{2.18}$$

o

$$z_0 < -z_\alpha \tag{2.19}$$

Donde α es el nivel de significancia (generalmente 0.05), y de la tabla de distrubución normal estándar se obtiene que $z_\alpha = z_{0,05} = 1{,}645$. En particular, el procedimiento para realizar las pruebas de significancia estadística aplicada a los métodos de sobre-muestreo se explica en el capítulo 5 de experimentos y resultados.

2.7.3. ANOVA (Análisis de Varianza)

En este caso, se investiga los diferentes efectos de los métodos de sobre-muestreo sobre el desempeño de un clasificador entrenado con clases que inicialmente están desbalanceadas. Los experimentos consistirán en aplicar varios métodos de sobre-muestreo varias veces sobre un conjunto de datos y probar su eficacia con respecto a un clasificador. La salida de este experimento se utilizará entonces para decidir qué método de sobre-muestreo obtiene mejores resultados. Hay que probar la siguiente hipótesis estadística.

- H_0: La varianza en los resultados de los métodos de sobre-muestreo es la misma.
- H_1: La varianza en los resultados de los métodos de sobre-muestreo es diferente.

H_0 es la hipótesis nula mientras que H_1 es la hipótesis alternativa.

Partimos de que se piensa que el método de sobre-muestreo influye en la mejora del desempeño de la clasificación. Se conoce como *factor* al método de sobre-muestreo utilizado y cada nivel será un método diferente. En este caso, llamamos *niveles del factor* a los métodos de sobre-muestreo. Se decide investigar a niveles de sobre-muestreo: método$_1$, método$_2$, ... , método$_a$. Se determina realizar n pruebas de validación cruzada con cada nivel de sobre-muestreo. A los *niveles del factor* también se les conoce como *tratamientos* y cada tratamiento tiene n observaciones o *réplicas*. El papel de la *aleatorización* es importante porque cada método de sobre-muestreo hace uso de ella. Entonces los resultados obtenidos en el desempeño podrían deberse a ella y por eso es importante hacer un análisis estadístico sobre los datos de un experimento con un sólo factor (el método de sobre-muestreo). En el análisis de varianza [20] los datos observados aparecerían como se muestra en la tabla 2.4.

Método de sobre-muestreo	Observaciones 1	2	...	n	Totales	Promedios
método$_1$	y_{11}	y_{12}	...	y_{1n}	y_1	$\bar{y}_1$
método$_2$	y_{21}	y_{22}	...	y_{2n}	y_2	$\bar{y}_2$
.	.	.	...	.	.	.
.	.	.	...	.	.	.
.	.	.	...	.	.	.
método$_a$	y_{a1}	y_{a2}	...	y_{an}	y_a	$\bar{y}_a$
					y	$\bar{y}$

Tabla 2.4: Datos típicos de un experimento con un sólo factor.

Un dato y_{ij} de la tabla representa la observación j-ésima hecha bajo el tratamiento i-ésimo. Estas observaciones pueden describirse como el *modelo estadístico lineal*.

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad i = 1, 2, ..., a \quad j = 1, 2, ..., a \tag{2.20}$$

donde Y_{ij} es una variable aleatoria que denota la observación ij-ésima, μ es un parámetro común a todos los tratamientos llamado *media global*, τ_i es un parámetro asociado con el tratamiento i-ésimo llamado *efecto del tratamiento i*-ésimo, y ε_{ij} es un componente del error aleatorio. A partir del desarrollo de esta ecuación se llega a las siguientes fórmulas para calcular las sumas de cuadrados del análisis de varianza con tamaños de las muestras iguales en cada tratamiento.

- Suma de cuadrados de los tatamientos:

$$SS_{Tratamientos} = n \sum_{i=1}^{a} (\bar{y}_i - \bar{y})^2 \tag{2.21}$$

- Suma de los cuadrados del error:

$$SS_E = SS_T - SS_{Tratamientos} \tag{2.22}$$

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	F_0
Tratamientos	$SS_{Tratamientos}$	$a-1$	$MS_{Tratamientos}$	$\frac{MS_{Tratamientos}}{MS_E}$
Error	SS_E	$a(n-1)$	MS_E	
Total	SS_T	$an-1$		

Tabla 2.5: El análisis de varianza para un experimento con un sólo factor.

- Suma total de cuadrados:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{y^2}{N} \tag{2.23}$$

- Cuadrado medio de los tratamientos:

$$MS_{Tratamientos} = SS_{Tratamientos}/(a-1) \tag{2.24}$$

- Cuadrado medio de error:

$$MS_E = \frac{SS_E}{a(n-1)} \tag{2.25}$$

donde N es el número de muestras tomandas ($a * n$). Los cálculos para este procedimiento de prueba se resumen en la tabla 2.5.

Una vez que se obtiene el valor de F_0, éste debe ser comparado con el *valor P* que es el nivel más bajo que llevaría al rechazo de la hipótesis nula H_0 con los datos dados. Si el valor F_0 es mayor que el *valor P* entonces se rechaza la hipótesis nula y se acepta la hipótesis alternativa, concluyendo que los resultados obtenidos no son producto del azar. El valor P es la probabilidad de que el estadístico de la prueba tome un valor que es al menos tan extremo como el valor observado del estadístico cuando la hipótesis nula H_0 es verdadera.

El análisis ANOVA para el desempeño de los métodos de sobre-muestreo se describe en el capítulo 5 de experimentos y resultados.

2.8. Resumen

En este capítulo se dieron las definiciones de los conceptos básicos que se involucran en este trabajo de investigación. Particularmente, se explicó el problema de clases desbalanceadas y su importancia, se presentaron las funciones de distancia que son utilizadas para determinar la vecindad entre instancias. Además se describió el proceso de normalización necesario para unificar el rango de valores que toman los atributos numéricos. Finalmente, se describió el Análisis de Varianza así como las pruebas de significancia estadística.

La información descrita en este capítulo es la base de los siguientes capítulos. La función de distancia euclidiana está involucrada en el capítulo 4, donde se describen los métodos de sobre-muestreo propuestos en esta tesis. Los clasificadores, las medidas de desempeño y los métodos de evaluación serán utilizados en el capítulo 5 de experimentos y resultados. En el siguiente capítulo se describen los métodos de sobre-muestreo que han sido propuestos anteriormente para intentar solucionar el problema de desbalance en las bases de datos.

Capítulo 3

Estado del Arte

Existen varios métodos que intentan mejorar el proceso de aprendizaje con datos cuyas clases están desbalanceadas. Estas soluciones pueden dividirse en dos categorías: a nivel algorítmico y a nivel de datos. Ambos tipos buscan favorecer el aprendizaje de la clase minoritaria. Los métodos que trabajan a nivel algorítmico modifican los algoritmos de minería de datos o implementan nuevos algoritmos para resolver el problema de clases desbalanceadas. Por otro lado, los métodos que trabajan a nivel de datos cambian la distribución del conjunto de datos desbalanceado, el proceso de aprendizaje de un clasificador se realiza con la distribución de datos balanceada. Cada grupo de métodos se describe a continuación.

3.1. Métodos a nivel algorítmico

Estos algoritmos buscan cambiar el sesgo en los algoritmos de aprendizaje para favorecer la clasificación de la clase minoritaria.

3.1.1. Aprendizaje sensible al costo

Una opción es incorporar métricas que consideren el problema de desbalance dentro de los algoritmos de aprendizaje y se conoce como aprendizaje sensible al costo. Esto es, se debe tener el costo asociado de hacer una mala clasificación y esa información se pueda representar mediante una matriz de costo. Se trata de introducir un costo mayor a los ejemplos mal clasificados de la clase minoritaria que a los de la mayoritaria. Dicha incorporación de costos sesga a los algoritmos de aprendizaje a cometer menos errores en la clase minoritaria.

Para explicar la importancia de asignar costos a una mala clasificación, consideremos el siguiente ejemplo: un médico oncólogo requiere clasificar células normales y células cancerígenas de los estudios realizados a sus pacientes. Si se clasifica una célula normal como cancerígena el costo de esa mala clasificación sería el monto de los gastos requeridos para confirmar tal diagnóstico. Por otro lado, si se clasificara una célula cancerígena como normal, el costo de tal error puede ser la muerte del paciente debido a que se diagnosticaría como sano cuando en realidad no lo es. En este ejemplo se ve cómo el costo de cometer un error en la clase de menor importancia (células sanas) no es tan grave como cometer un error en la clasificación de la clase más importante (células cancerígenas).

Una de las teorías de costo (riesgo) más populares adoptadas para medir el costo en proyectos es la Teoría Estadística Bayesiana [9], nombrada así en honor a Thomas Bayes, un matemático británico del siglo XVIII. La Teoría Bayesiana se basa en la enumeración de diferentes eventos posibles y la asociación de cada uno con una probabilidad de ocurrencia. Por medio de la cuantificación del impacto de cada evento, y la multiplicación por su correspondiente probabilidad de ocurrencia, se pueden calcular los “daños esperados” (costos) de cada factor de riesgo.

Suponiendo que el clasificador proporciona la probabilidad de cada clase, al momento de hacer la evaluación del algoritmo, se toma en cuenta la matriz de costo para predecir la clase que tenga el menor costo de error, que se puede obtener al multiplicar la matriz de costo por las probabilidades asociadas por el clasificador a cada clase. Un ejemplo de esta incorporación de costos es cambiando el umbral de decisión [17] de los métodos probabilistas como el clasificador bayesiano simple.

MetaCost

Hay otra posibilidad de utilizar costos sin hacer modificación del algoritmo de clasificación y utilizar los costos fuera de él, esto es lo que hace el algoritmo MetaCost [7]. Este algoritmo emplea una combinación de clasificadores, basada en la generación de clasificadores de varias muestras de los ejemplos (*bagging - Bootstrap Aggregating*), para reetiquetar los ejemplos de entrenamiento. Las estimaciones de probabilidad de las clases se realizan aprendiendo clasificadores múltiples y, por cada ejemplo, usa la fracción del voto total para cada clase como estimación de su probabilidad. Después se reetiqueta cada ejemplo de entrenamiento con la clase que minimice el costo. Este conjunto reetiquetado se emplea para aprender un nuevo clasificador sensible al costo.

Toma directa de decisiones sensible al costo

En [33] se presenta un algoritmo de meta-aprendizaje (*Direct Cost-Sensitive Decision Making*), en el que se asume que no todos los costos son conocidos ni que el costo es el mismo para todos los ejemplos como se hace en MetaCost. Este es un caso especial en el que para estimar el costo de error de cada ejemplo se utilizan la técnica de regresión lineal múltiple por mínimos cuadrados a partir de ciertos costos conocidos. El algoritmo emplea también una combinación de clasificadores pero basada en

la generación secuencial de clasificadores (*boosting*). Este algoritmo altera la distribución original de los ejemplos de la siguiente forma: el número de veces que se repite un ejemplo viene dado por un factor proporcional al costo relativo de cada ejemplo y luego se aplica a cada ejemplo un criterio de pertenencia para formar parte del conjunto de ejemplos de entrenamiento. El problema de este algoritmo es que no es determinista la elección del criterio de aceptación de ejemplos de entrenamiento.

Problema del aprendizaje sensible al costo

El problema principal con el aprendizaje sensible al costo es que estos costos son generalmente desconocidos y difíciles de encontrar porque dependen del dominio en cuestión. Para algunos dominios será fácil encontrar este costo pero para otros no. Sin embargo, existe una relación directa entre aumentar el costo de clasificación y aumentar el número de ejemplos de la clase minoritaria [8].

Otros autores también sugieren que con técnicas de re-muestreo se pueden generar clasificadores similares a aquellos producidos cuando se varía el umbral de decisión o la matriz de costos (que contiene los costos asociados a una mala clasificación), por ejemplo en [17], pero sin necesidad de conocer estos últimos.

3.1.2. Algoritmos específicos

Se han desarrollado también algoritmos específicos para clases desbalanceadas. La estrategia es no utilizar los algoritmos de aprendizaje existentes. La razón principal es que los casos raros pueden ser difíciles de representar y por lo mismo de generar, ya que pocos datos ofrecen poca guía al algoritmo de aprendizaje y los métodos de aprendizaje en general no consideran eso. Hacer algún tipo de búsqueda más completa entre los métodos de aprendizaje útiles para clases desbalanceadas es, en general,

computacionalmente muy costoso, aunque se han hecho algunas propuestas como en [22] y algunas otras basadas en algoritmos genéticos [3].

Otra estrategia consiste en hacer combinaciones de algoritmos para mejorar los resultados. Por ejemplo se ha utilizado C4.5 [21], como una primera pasada, y en las ramas con pocos ejemplos (*small disjuncts*) se usa otro tipo de clasificador como uno basado en instancias [26] o uno basado en algoritmos genéticos [3].

3.1.3. Algoritmos para una sola clase

Existen algunas propuestas de algoritmos de aprendizaje de una sola clase. La idea principal es generar sólo un clasificador para la clase minoritaria, el cual en general es menos susceptible de sufrir un sesgo por la clase mayoritaria (por ejemplo en [22]).

El algoritmo *Rule Extraction for MEdical Diagnostic* (REMED) [18], actualmente genera una sola regla para una sola clase, buscando principalmente entendimiento de los modelos sobre dominios médicos.

El inconveniente de estos métodos es que, en general, su desempeño puede variar al ser aplicados a problemas particulares.

3.2. Métodos a nivel de datos

En este nivel se han propuesto también varias formas de re-muestreo [5], que pueden ser divididas a su vez en métodos de sub-muestreo y métodos de sobre-muestreo. Ambos grupos buscan que la distribución de datos de las clases sea similar.

- Sub-muestreo: Selecciona un conjunto de instancias de la clase mayoritaria y éstas son eliminadas con el propósito de disminuir el tamaño de la clase mayoritaria.
- Sobre-muestreo: Trabaja con las instancias de la clase minoritaria para generar nuevas instancias y así incrementar el tamaño de esta clase.

Como se mencionó anteriormente, re-muestrear un conjunto de datos es equivalente a evaluar un clasificador con diferentes umbrales de decisión.

3.2.1. Métodos de Sub-Muestreo

Estos algoritmos reducen el tamaño de la clase mayoritaria buscando igualarlo al de la clase minoritaria.

Sub-muestreo aleatorio

RUS (por sus siglas en inglés: *Random Under Sampling*) es el más simple de los métodos de sub-muestreo. Consiste en seleccionar aleatoriamente una cantidad de instancias de la clase mayoritaria para ser eliminadas del conjunto. La principal desventaja de este método es que puede eliminar información útil para la clasificación. Sin emabargo, en [28] se muestra que el desempeño de RUS es superior en varias ocasiones que otros métodos de sub-muestreo.

Tomek Link

Este método [27] puede ser definido como sigue: Dadas dos instancias E_i y E_j de clases opuestas, sea $d(E_i, E_j)$ la distancia entre ellas. Un par $A(E_i, E_j)$ se llama *Tomek*

link si no hay una instancia E_l tal que $d(E_i, E_l) < d(Ei, Ej)$ o $d(E_j, E_l) < d(E_i, E_j)$. Si dos instancias forman un *Tomek link*, entonces una de ellas es ruido o ambas están en el borde de las clases. Como método de sub-muestreo, sólo las instancias *Tomek link* de la clase mayoritaria son eliminadas.

Regla Consensada del Vecino más Cercano

La regla CNN (*Condensed Nearest Neighbor Rule*) [6] se utiliza para encontrar un subconjunto consistente de instancias, que es un subconjunto del conjunto de entrenamiento T, que debe ser suficiente para clasificar todo T utilizando sólo un vecino más cercano. Se basa en la construcción, paso a paso, de un nuevo conjunto de objetos V, moviendo hacia él cada elemento de la matriz de entrenamiento T si éste es erróneamente clasificado por los objetos que ya están en V. El proceso se repite hasta que se eliminen todos las inconsistencias (errores).

Aunque, como su mismo autor reconoce, usualmente este método no obtiene el subconjunto mínimo consistente, pudiendo quedar, dependiendo del orden de los objetos, objetos muy alejados de la frontera. Por ejemplo, el primer objeto siempre es adicionado al resultado.

Selección de un solo lado

Método propuesto por Kubat y Matwin [15] (*One-sided selection*). Esta estrategia consiste en seleccionar un subconjunto representativo de las instancias de la clase mayoritaria. El requerimiento es que el clasificador mantenga todas las instancias minoritarias y elimine sólo las mayoritarias (usando CNN). Se utilizan cuatro heurísticas para detectar las instancias menos confiables de la clase mayoritaria.

1. Instancias que presentan ruido en cuanto a la etiqueta de la clase.
2. Instancias en los bordes de las clases.
3. Instancias que son redundantes.
4. Instancias *seguras* que serán parte para futura clasificación.

El inconveniente es que no hay una forma clara de definir qué instancias son en realidad ruido.

Regla de Limpieza de vecinos

NCL (*Neighborhood Cleaning Rule*) [16] utiliza ENN (*Wilson's Edited Nearest Neighbor Rule*) [30] para eliminar instancias de la clase mayoritaria. ENN elimina toda instancia cuya etiqueta de clase difiera de la clase de a lo menos dos de sus tres vecinos más cercanos. Para problemas con dos clases el algoritmo se puede describir de la siguiente manera: Para cada instancia E_i en el conjunto de entrenamiento, se calculan sus tres vecinos más cercanos. Si E_i pertenece a la clase mayoritaria y la clasificación dada para sus tres vecinos contradice la clase original de E_i, entonces E_i es eliminado. Si E_i pertenece a la clase minoritaria y sus tres vecinos clasifican equivocadamente E_i, entonces los vecinos más cercanos a E_i que pertenecen a la clase mayoritaria son eliminados. Este método frecuentemente es apropiado pero el costo computacional en cuanto almacenamiento y procesamiento es muy alto y se presentan restricciones cuando el conjunto de entrenamiento es grande.

3.2.2. Métodos de Sobre-Muestreo

Los métodos de sobre-muestreo aumentan el tamaño de la clase minoritaria con el fin de obtener un conjunto de datos balanceado. El sobre-muestreo aleatorio es un método

muy simple, mientras que SMOTE junto con las dos variantes de Borderline-SMOTE son métodos que buscan crear nuevas instancias de una manera “inteligente”.

Sobre-muestreo aleatorio

ROS (por sus siglas en inglés: *Random Over-Sampling*) es el más simple de los métodos de sobre-muestreo. Selecciona aleatoriamente un grupo de instancias pertenecientes a la clase minoritaria y éstas son replicadas cierta cantidad de veces, de acuerdo al porcentaje de sobre-muestreo que vaya a realizarse. Sin embargo, este método puede hacer la región de decisión más pequeña y específica para el clasificador, lo que causará sobre ajuste. Pero, al igual que RUS en [28], se muestra que su desempeño es mejor en varias ocasiones que otros métodos de sobre-muestreo.

SMOTE

SMOTE (por sis siglas en inglés *Synthetic Minority Over-sampling Technique*) [4] propone un enfoque de sobre-muestreo en el que la clase minoritaria es sobre-muestreada por la creación de instancias sintéticas en lugar de sólo sobre-muestreo por replicación. Su idea principal es formar nuevas instancias interpolando entre pares de instancias cercanas entre sí de la clase minoritaria. El algoritmo SMOTE se muestra en la tabla 3.1. Vea la figura 3.1 donde se muestra la generación de una instancia sintética. Gráficamente, si pensamos en un dominio con dos atributos numéricos, podemos ver que la instancia sintética se genera en la región del rectángulo definido por la instancia tomada como punto de partida y el vecino seleccionado.

SMOTE es un algoritmo muy aceptado por la comunidad científica. Sin embargo, hay aspectos que deben ser considerados para mejorar su desempeño. Primero, la forma inteligente de generar nuevos valores sólo aplica a atributos numéricos.

1. Recibe como parámetro el porcentaje de ejemplos a sobre-muestrear.
2. Calcula el número de ejemplos que tiene que generar.
3. Calcula los k vecinos más cercanos (por defecto $k = 5$) de cada ejemplo de la clase minoritaria.
4. Genera los ejemplos siguiendo este proceso:
   4.1 Para cada ejemplo de la clase minoritaria, elige aleatoriamente el vecino a utilizar para crear el nuevo ejemplo.
   4.2 Para cada atributo del ejemplo a sobre-muestrear, calcula la diferencia entre el vector de atributos muestra y el vecino elegido.
   4.3 Multiplica esta diferencia por un número aleatorio entre 0 y 1.
   4.4 Suma este último valor al valor original de la muestra.
5. Regresa el conjunto de ejemplos sintéticos.

Tabla 3.1: Algoritmo SMOTE

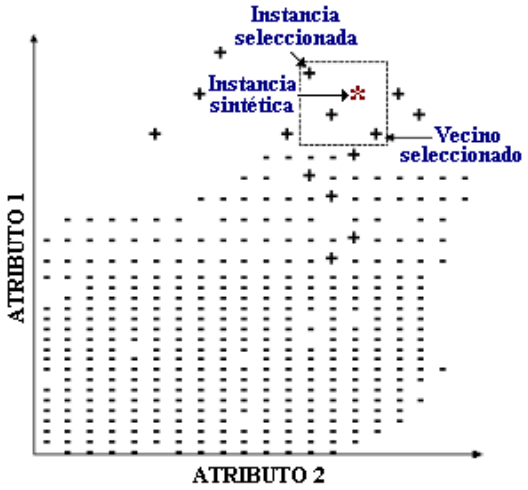

Figura 3.1: Generación de una instancia sintética utilizando SMOTE. Selecciona una instancia de la clase minoritaria y aleatoriamente escoge a uno de sus k vecinos más cercanos de la misma clase. En el espacio que que se encuentra entre las dos instancias se genera la instancia sintética interpolando los valores de los atributos de ambas instancias.

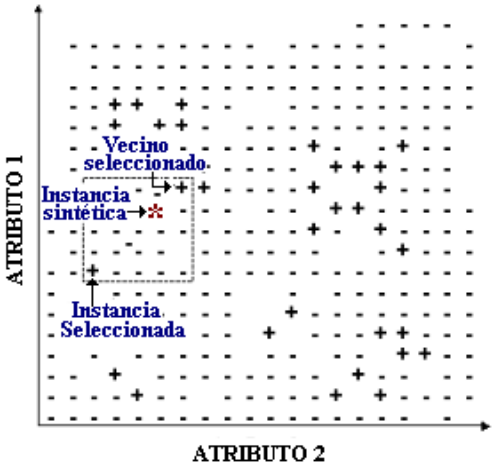

Figura 3.2: Generación con SMOTE de una instancia sintética en un área de la clase mayoritaria. Al seleccionar una instancia de la clase minoritaria puede suceder que entre ella y su vecino más cercano de la misma clase existan instancias de la clase mayoritaria, lo que ocasionará traslape de clases.

Para el caso de atributos nominales se considera el voto mayoritario de los k-vecinos más cercanos. Pero si el porcentaje de sobre-muestreo a realizar es muy alto las instancias se generarán con el mismo valor nominal en todos los casos. Segundo, pueden existir instancias aisladas. En este caso al seleccionar a uno de los vecinos más cercanos éste estará demasiado alejado y al interpolar entre ambas instancias se generará una instancia en una región que seguramente sea de la clase mayoritaria. Tercero, similar al caso anterior, si se selecciona aleatoriamente uno de los k-vecinos más cercanos podría suceder que éste sea también muy lejano y al interpolar se genere la instancia en un área inapropiada y se genere así un modelo incorrecto de los datos (Figura 3.2). Por último, no tiene una forma definida de elegir el valor de k. Aunque generalmente este valor se establece en 5 se hace de manera arbitraria, ya que debería depender del dominio en cuestión.

Borderline-SMOTE

El método Borderline-SMOTE [13] es un método basado en SMOTE [4] con la diferencia de que sólo aquellas instancias que se encuentran en el borde de las clases son las que se sobre muestrean. Este método tiene dos versiones: Borderline-SMOTE1 y Borderline-SMOTE2, que se describen a continuación.

Primero se determinan aquellas instancias que se encuentran en el borde de la clase positiva (o clase minoritaria); después, las instancias sintéticas se generan de esas instancias del borde y agregadas al conjunto de entrenamiento original. Suponga que el conjunto original completo es T, la clase positiva (minoritaria) es P y la negativa (mayoritaria) es N, y

$$P = \{p_1, p_2, ..., p_{pnum}\}, N = \{n_1, n_2, ..., n_{nnum}\} \tag{3.1}$$

donde $pnum$ y $nnum$ son el número de instancias de la clase positiva y la negativa respectivamente. El procedimiento detallado de Borderline-SMOTE1 es como sigue:

1. Para cada $p_i (i = 1, 2, ..., pnum)$ en la clase positiva P, se calculan sus m vecinos más cercanos del conjunto de entrenamiento completo T. El número de ejemplos negativos entre los m vecinos más cercanos se denota por $m'(0 < m' < m)$.

2. Si $m' = m$, es decir, todos los vecinos más cercanos de p_i son instancias negativas, p_i es considerada ruido y no se utiliza en los siguientes pasos. Si $m/2 <= m' < m$, esto es, el número de vecinos negativos más cercanos de p_i es mayor que el número de positivos, p_i se considera en el borde de la clase y se coloca en un conjunto llamado $DANGER$. Si $0 <= m' < m/2$, p_i se considera segura y no se considera en los siguientes pasos.

3. Las instancias en $DANGER$ son las que se encuentran en el borde de la clase

positiva, y se puede ver que $DANGER$ está contenido en P. Entonces

$$DANGER = \{p'_1, p'_2, ..., p'_{dnum}\}, 0 \leq dnum \leq pnum \tag{3.2}$$

Para cada instancia en $DANGER$, se calculan sus k vecinos más cercanos en P.

4. Se generan $s * dnum$ instancias positivas sintéticas de los datos en $DANGER$, donde s es un número entero entre 1 y k. Para cada p'_i, se seleccionan aleatoriamente s vecinos más cercanos de sus k vecinos más cercanos en P. Primero, se calculan las distancias, $dif_j (j = 1, 2, ..., s)$ entre p'_i y sus vecinos más cercanos en P, después se multiplica dif_j por un número aleatorio $r_j (j = 1, 2, ..., s)$ entre 0 y 1, finalmente, son generadas s nuevas instancias sintéticas entre p'_i y sus vecinos más cercanos.

$$nueva\_instancia_j = p'_i + r_j * dif_j \quad j = 1, 2, ..., s \tag{3.3}$$

El procedimiento anterior se repite para cada p'_i en $DANGER$ y puede lograr $s * dnum$ instancias sintéticas. Este paso es similar a SMOTE [4].

En el procedimiento de arriba, p_i, n_i, p'_i, dif_j y $nueva\_instancia_j$ son vectores. Las instancias sintéticas nuevas se generan a lo largo de la línea entre las instancias del borde de la clase positiva y sus vecinos más cercanos de la misma clase.

Borderline-SMOTE2 no sólo genera instancias sintéticas para cada ejemplo en $DANGER$ y sus vecinos más cercanos en la clase positiva P, también lo hace de sus ejemplos más cercanos de la clase negativa N. La diferencia es que la instancia y su vecino se multiplica por un número aleatorio entre 0 y 0.5, de esta manera las instancias generadas estás más cerca a la clase positiva que de la negativa.

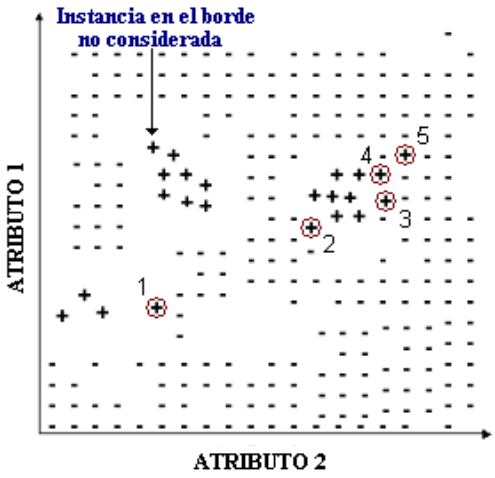

Figura 3.3: Inconveniente de Borderline-SMOTE. No es capaz de encontrar todas las instancias que definen los bordes (encerrados en un círculo). Una instancia sintética generada entre bordes de diferentes regiones puede caer en una zona de la clase negativa.

El principal inconveniente de estos métodos, además de los mencionados para SMOTE, es que no siempre son capaces de encontrar todos los bordes de las clases (cuando están muy separadas las clases). Por ejemplo, vea la figura 3.3, en ella las instancias detectadas como borde por Borderline-SMOTE están marcadas con un círculo y enumeradas del 1 al 5. Estas instancias están en el conjunto *DANGER* porque la mayoría se sus $k = 5$ vecinos más cercanos pertenecen a la clase mayoritaria. Por ejemplo, si observamos las cinco vecinas más cercanas de la instancia número 1, todas ellas pertenecen a la clase mayoritaria. Aunque hay otras instancias que, gráficamente, se muestra que están también en el borde, éstas no son detectadas por Borderline-SMOTE porque la mayoría de sus vecinos más cercanos, o todos, son de la clase minoritaria. Pueden presentarse casos en las bases de datos donde la separación entre instancias de diferentes clases sea lo suficientemente grande como para no detectar elementos del borde. Entonces, al no tener instancias en el conjunto $DANGER$ no se lleva a cabo el proceso de sobre-muestreo. Este método trabaja muy bien sólo cuando los bordes de las clases están bien definidos.

3.3. Resumen

Este capítulo dió una descripción de los diferentes enfoques que se han tomado para solucionar el problema de clases desbalanceadas. Primero se presentaron los métodos a nivel de algoritmo que incluyen modificación de costos o creación de nuevos algoritmos. Después se fueron presentados métodos a nivel de datos que cambian la distribución de instancias con sobre o sub muestreo.

Considerando los aspectos que deben ser reforzados en los métodos de sobremuestreo, en el siguiente capítulo se proponen dos nuevos métodos para generación inteligente de instancias sintéticas.

Capítulo 4

Métodos de Sobre-Muestreo propuestos

Como se describió en el capítulo 3, el principal problema del sobre muestreo aleatorio es que tiende a sobre ajustar el clasificador con la clase minoritaria. Esto significa que el clasificador aprenderá muy bien de las instancias de entrenamiento pero al clasificar nuevas instancias éstas pueden no ser clasificadas correctamente. Esto se debe a que con este tipo de sobre muestreo el área de la clase minoritaria se vuelve más específica. En el caso de los métodos inteligentes, éstos toman pares de instancias de la clase minoritaria y generan nuevas instancias entre ellas interpolando los valores de los atributos numéricos. Aunque en general estos métodos han reportado resultados satisfactorios, en el sentido de aumentar el porcentaje de instancias correctamente clasificadas de la clase minoritaria, aún hay aspectos que deben mejorarse.

En el caso de SMOTE pueden existir instancias aisladas o traslape de clases. En tal situación, al considerar los 5 vecinos más cercanos y elegir uno puede suceder que al generar una instancia entre ellos, éste caiga en una zona dominada por la clase mayoritaria, en cuyo caso la clasificación de la clase mayoritaria se verá afectada

como se ve en la figura 3.2 del capítulo anterior. Además, considera $k = 5$ vecinos más cercanos pero este es un valor que depende del problema. Sería más apropiado considerar las características éste, como son la distribución de instancias, empalme de clases, entre otras. Por último, para asignar un valor a un atributo nominal se utiliza el voto mayoritario de los vecinos más cercanos. Esto es, se verifica qué valor de ese atributo es el que se repite más entre los k vecinos más cercanos y ese se asigna a la nueva instancia. Sin embargo, el uso de esta técnica presenta un inconveniente cuando el porcentaje de sobre-muestreo es muy alto, ya que todas las instancias sintéticas que se generen a partir de una instancia dada tendrán el mismo valor (el que más se repite), y esto puede provocar sobre ajuste de ese atributo. Borderline-SMOTE sólo obtiene buenos resultados cuando los bordes de las clases están bien definidos.

A partir de lo anteriormente expuesto, se proponen dos nuevas técnicas de generación inteligente de instancias. La idea principal es crear instancias sintéticas en áreas dominadas por la clase minoritaria y de manera local en lugar de global. Al considerar áreas para las instancias minoritarias no será necesario establecer un valor k para los vecinos más cercanos, ya que todas las instancias dentro de una misma región serán vecinas entre ellas. Para tratar el caso especial de los atributos nominales se propone hacerlo probabilísticamente de acuerdo a la distribución de valores.

4.1. Generación de grupos

El método propuesto para definir las áreas correspondientes a la clase minoritaria es agrupar las instancias de tal manera que entre ellas sean vecinas (más adelante se verá que pueden incluirse algunas instancias mayoritarias siempre y cuando no rebasen en número a las minoritarias). Esto se puede ver como la generación de grupos de instancias minoritarias. Se trata de definir las áreas que dominan las instancias de la clase minoritaria. De esta manera, cada grupo contendrá instancias minoritarias cuya distan-

cia entre ellas sea pequeña comparada con la distancia que tengan con las instancias de las instancias mayoritarias.

Para llevar a cabo el procedimiento de generación de grupos es necesario conocer las distancias entre cada par de instancias (sin importar a qué clase pertenezcan). Para calcular las distancias entre instancias, en esta tesis se eligió la medida $HVDM$ (*Heterogeneous Value Difference Metric*) que contempla atributos tanto numéricos como nominales (Sección 2.2). $HVDM$ utiliza una función $euclidiana$ para atributos numéricos y vdm para nominales. Esto es para definir la cercanía entre instancias y así llevar a cabo la formación de los grupos con instancias minoritarias que estén muy cercanas entre ellas. Por ello se genera una matriz cuadrada M de n filas por n columnas, donde n es el número total de instancias y el valor M_{ij} de cada celda es la distancia entre la instancia i y la instancia j. M es una matriz simétrica ya que $M_{ij} = M_{ji}$ porque la distancia de i a j es igual a la distancia j a i. Cada vez que se requiera la distancia entre dos instancias se obtendrá de esta matriz. Contar con una matriz de distancia es opcional, hay que considerar el tamaño del conjunto de datos, ya que si éste es muy grande será preferible calcular la distancia entre pares de instancias cada vez que se requiera para evitar generar una matriz tan grande.

En principio se generarán T grupos, uno por cada elemento de la clase a sobre muestrear, ya que cada instancia se tomará como semilla para iniciar un nuevo grupo. Para cada semilla se construye una lista ordenada ascendentemente (de acuerdo a la distancia) de todas las demás instancias (sin importar la clase a la que pertenece). A partir de esta lista se incluyen sólo las instancias minoritarias que aparezcan antes de una instancia mayoritaria. Esto se ilustra en la figura 4.1, cada instancia se identifica por la letra I acompañada de un número y delante de ese identificador se indica con un signo “+” o un signo “-” a qué clase pertenece (minoritaria o mayoritaria). De esta lista se toman de izquierda a derecha todas las instancias minoritarias que aparezcan antes de la primera instancia mayoritaria, éstas serán agregadas al grupo correspon-

ID instancia	I1 +	I2 +	I3 +	I4 +	I5 +	I6 -	I7 -	I8 -	...	IN +
Distancia	0.50	0.62	0.65	0.71	0.78	0.86	0.88	0.93	...	1.94

corte

Figura 4.1: Lista de instancias ordenada ascendentemente de acuerdo a su distancia con la instancia semilla.

diente a la semilla tomada. Entonces, en el ejemplo de la figura 4.1, el grupo formado contendrá (además de la instancia semilla) las instancias I1, I2, I3, I4 e I5.

Una vez que se han formado todos los grupos a partir de cada semilla (cada instancia del conjunto) se verifica cuáles están contenidos uno dentro de otro. Por ejemplo, si el grupo A contiene a las instancias I1, I2, I3, I4 e I5 y el grupo B contiene a las instancias I2, I3 e I4 entonces el grupo B está contenido en el grupo A, por lo tanto únicamente se conserva el grupo A y el B es descartado.

Finalmente, se verifica cada grupo con todos los demás con el propósito de verificar si tienen algunos elementos en común. Si la intersección de ellos no es vacía entonces se intenta unirlos en un solo grupo que puede contener instancias de la clase mayoritaria pero sin que ellas superen en número a las de la clase minoritaria. El procedimiento para la generación de todos los grupos se describe en la tabla 4.1.

Como se menciona en el paso 3 de la tabla 4.1, una vez obtenidos los grupos se verificará la intersección entre cada par de ellos, si la intersección es vacía se quedan como están pero sí hay elementos en ella (traslape de grupos) entonces intentará fusionar ese par de grupos como se describe en la tabla 4.2. La fusión de grupos se da porque, dados dos grupos, éstos podrían estar separados sólo por unas cuantas instancias mayoritarias. Entonces el área que abarcan ambos grupos en total puede ser considerada como área de la clase minoritaria aún cuando existan algunas instancias

1. Para cada instancia de la clase minoritaria:
   1.1 Crear un nuevo grupo cuyo primer elemento sea la instancia elegida como semilla.
   1.2 Crear una lista ordenada ascendentemente con todas las demás instancias (de ambas clases) de acuerdo a su distancia con la instancia semilla.
   1.3 Si el primer elemento de la lista ordenada es de otra clase entonces el grupo sólo se quedará con un elemento (la semilla), porque el ejemplo más cercano a la semilla es de otra clase. Se trata entonces de un ejemplo aislado.
   Sino entonces se agregan todos los elementos de la misma clase que la semilla que aparecen antes de uno de clase contraria.
2. Fusionar pares de grupos donde uno de ellos esté contenido en el otro.
(vea la figura 4.2)
3. Fusionar grupos con elementos comúnes sin que se incluyan más ejemplos de la clase mayoritaria que de la minoritaria.

Tabla 4.1: Algoritmo de generación de grupos de instancias minoritarias.

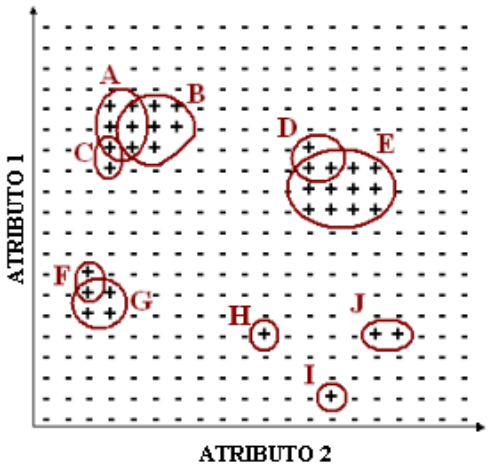

Figura 4.2: Conjunto de grupos después de fusionar aquéllos que estaban contenidos unos en otros.

1. Eliminar del grupo más pequeño las instancias que están en la intersección. 2. Una vez que los grupos son disjuntos se obtienen las instancias extremas, que son dos instancias cuya distancia entre ellas es la máxima de todas las distancias entre pares de instancias (considerando ambos grupos). 3. Obtener un punto medio o *centroide* entre las instancias extremas. Este centroide será una nueva instancia (instancia media). Para obtenerlo, en el caso de atributos lineales se obtiene el promedio entre los valores de los atributos de las instancias extremas. Para atributos nominales: si ambas instancias tienen el mismo valor entonces éste se conserva, si son diferentes entonces el valor que se asignará a la instancia media será un vector de dos elementos que contendrá los valores de ese atributo en cada instancia. Ver figura 4.3. 4. La distancia de la instancia media a las instancias extremas es el radio, con el que se traza una circunferencia que contempla instancias de ambas clases. Se cuenta el total de instancias de cada clase, si el número de instancias de la clase de interés (minoritaria) es mayor que el de la clase contraria entonces se fusionan los grupos (figura 4.4), de lo contrario se quedan como están.

Tabla 4.2: Algoritmo para fusionar grupos con traslape entre ellos.

mayoritarias. Al fusionar estos grupos estamos dando mayor amplitud a la región donde se generarán las instancias sintéticas.

Con el procedimiento de la tabla 4.2 se busca generar grupos con el mayor tamaño posible y definir áreas mayormente influenciadas por la clase minoritaria. Aunque se puedan incluir algunas instancias mayoritarias, se asumirá que hay sólo instancias minoritarias y se podrá realizar el sobre-muestreo en esa área. A partir de estos grupos se puede aplicar cualquiera de los dos métodos propuestos en las siguientes secciones. El primero, llamado GIS-G, genera instancias como SMOTE pero sólo interpola entre instancias de un mismo grupo. El segundo, llamado GIS-GF, crea valores numéricos considerando la desviación estándar de los datos y sólo requiere de una instancia como semilla para crear una instancia sintética. Ambos métodos presentan una forma inteligente de generar valores nominales considerando la distribución de valores en los datos y asignando los nuevos valores de manera probabilista.

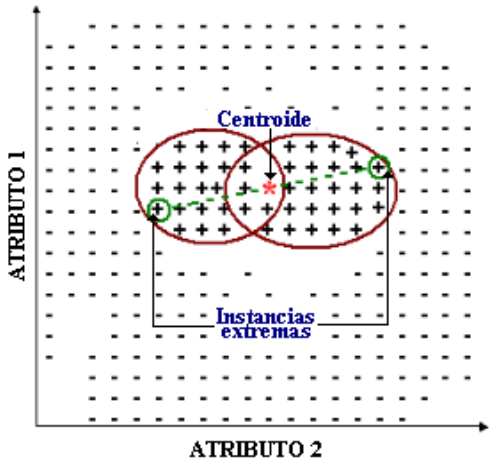

Figura 4.3: Centroide de instancias extremas de dos grupos.

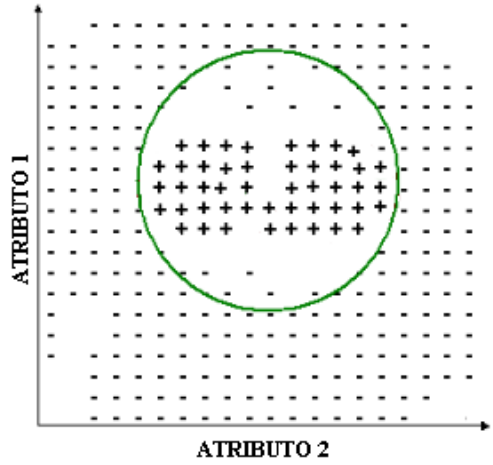

Figura 4.4: Fusión de grupos.

4.2. GIS-G (Generación de Instancias Sintéticas mediante formación de Grupos)

Dado un conjunto de grupos tal que n_{grupos} es el total de ellos y n_i es el tamaño del i-ésimo grupo, para generar N instancias sintéticas, primero se verifica el número total de instancias a generar en cada grupo que debe ser proporcional al tamaño de éste. Después, recorriendo cíclicamente las instancias del grupo hasta completar el total de instancias a generar, se selecciona una instancia del grupo para tomar como semilla. A continuación se selecciona aleatoriamente una instancia del mismo grupo (que no sea la semilla) y con ella se efectúa el proceso de interpolación para atributos numéricos similar a SMOTE para crear una instancia sintética. Para el caso de atributos nominales se considera la distribución de los valores en el grupo como se describe a continuación:

Supongamos que un atributo nominal puede tomar uno de tres valores diferentes: "rojo", "azul" o "amarillo". Y supongamos también que hay 10 instancias en el grupo con la siguiente distribución de valores en para ese atributo: 5 con valor "rojo", 3 con valor"azul" y 2 con valor "amarillo". Si hubiera que generar n instancias, SMOTE o Borderline-SMOTE las generarían todas con el valor "rojo" (por voto mayoritario). En este caso se toma la distribución de valores como probabilidades, esto es, 50 % de probabilidad de asignar "rojo", 30 % de probabilidad de asignar "azul" y 20 % de probabilidad de asignar "amarillo". Se debe generar un valor aleatorio x en $[0, 1)$, ese intervalo se divide de acuerdo a los porcentajes de valores. Así, si el valor $x < 0{,}5$ entonces el valor asignado será "rojo", si $0{,}5 \leq x < 0{,}8$ el valor será "azul" y si $0{,}8 \leq x$ el valor será "amarillo".

El procedimiento de GIS-G es el que se describe en la tabla 4.3 y, como se muestra en la figura 4.5, con él es posible generar instancias en áreas mayormente dominadas por la clase minoritaria, de esta manera se evita generarlas en áreas de la clase mayoritaria. Si bien es cierto que la clase mayoritaria es la de menor interés,

1. Para cada grupo, si el tamaño del grupo es mayor que 1:
   1.1 Calcular el número de instancias a generar en ese grupo.
   $N_i = (n_i/N) * 100$
   1.2 Generar N_i instancias en ese grupo recorriendo cíclicamente las instancias del grupo y tomando cada instancia como referencia para crear una instancia sintética como sigue:
      1.2.1 Dada la instancia de referencia, seleccionar aleatoriamente una de las instancias que pertenecen al mismo grupo y generar con ella una instancia sintética como sigue:
      1.2.2 Para cada atributo:
      1.2.2.1 Si el tipo de atributo es numérico entonces el valor de ese atributo de la instancia sintética se calcula interpolando los valores de los atributos del par de instancias seleccionadas.
      1.2.2.2 Si el tipo de atributo es nominal entonces se verifica la proporción de valores presentes en el grupo. El porcentaje de cada valor en el grupo se tomará como una probabilidad y de acuerdo a ella se asignará el valor a la instancia sintética.

Tabla 4.3: Algoritmo GIS-G.

también es cierto que debe evitarse perjudicar su clasificación en la mayor medida que sea posible. Además, también presenta las siguientes ventajas:

- Al generar las instancias sintéticas entre elementos de un mismo grupo, ya no es necesario establecer el parámetro k que indica el número de vecinos a considerar.

- Para el caso de atributos de tipo nominal, presenta una ventaja sobre SMOTE y Borderline-SMOTE en cuanto a que no sólo genera los valores de estos atributos replicando los que predominan, sino que se realiza de una manera más equitativa de acuerdo a la proporción de valores que hay en cada grupo como en el ejemplo que se mostró.

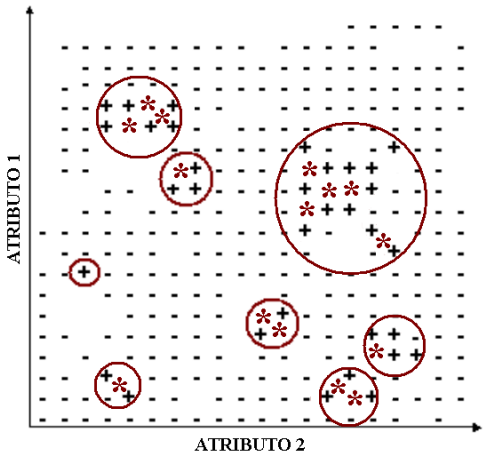

Figura 4.5: Generación de instancias dentro de cada grupo de la clase minoritaria, las instancias sintéticas están representadas por un asterisco.

4.3. GIS-GF (Generación de Instancias Sintéticas mediante formación de Grupos y Fluctuaciones)

En algunos casos al generar grupos, dada la distribución de las instancias, es posible que los grupos contengan pocas instancias e incluso sólo una sin que esto necesariamente quiera decir que se trata de instancias aisladas. Puede ser que sean instancias cuyos vecinos más cercanos sean de la clase mayoritaria aunque el área en la que se encuentran sea de la clase minoritaria. Esto se aprecia mejor en un ejemplo como el que se muestra en la figura 4.6. Si se toma cada instancia como punto de partida para formar un nuevo grupo y se calculan los vecinos más cercanos para determinar los demás elementos del grupo, estos vecinos serán de la clase mayoritaria. El grupo quedará con un sólo elemento sin que esto signifique que se trata una instancia aislada ya que, como se observa en la figura, el área de la clase minoritaria es la parte superior. Con los métodos presentados, para generar una instancia sintética siempre es necesario contar con un par de instancias para poder interpolar los valores de sus atributos.

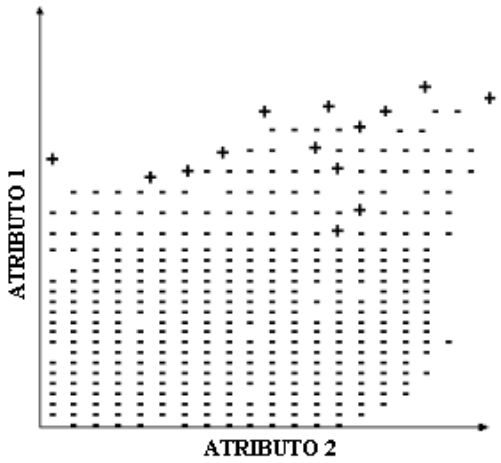

Figura 4.6: Instancias de la clase minoritaria que formarían grupos con un único elemento. Si se consideran los vecinos más cercanos, éstos pertenecen a la clase mayoritaria y el grupo se quedará con un solo elemento, aún cuando el área en la que se encuentra la instancia es de la clase minoritaria.

El método GIS-GF propone generar instancias a partir de sólo una instancia de la clase minoritaria y hacerlo "alrededor" de ella (en un sentido gráfico). El término *fluctuaciones* se emplea para referirnos a los movimientos de salto (aumento o disminución de valores) que se realizan para asignar valor a un atributo numérico.

La desviación estándar se define como la raíz cuadrada de los cuadrados de las desviaciones de los valores de la variable respecto a su media. Mientras que la desviación media obtiene el promedio de los valores absolutos de las desviaciones de los valores respecto de la media, también es conocida como promedio de desviación. El uso de estas desviaciones permite generar valores a partir de sólo una instancia controlando el área donde se generará la instancia sintética. Se utiliza la menor de ellas para no generar nuevas instancias muy alejadas de la original. Tal control está dado en términos de los valores que pueden tomar los datos. El procedimiento GIS-GF se describe en la tabla 4.4.

1. Para cada grupo:
   1.1 Calcular el número de instancias a generar en ese grupo.
   $N_i = (n_i/N) * 100$
   1.2 Generar N_i instancias en ese grupo recorriendo cíclicamente las instancias del grupo y tomando cada instancia como referencia para crear una instancia sintética como sigue:
   1.2.1 Para cada atributo:
   1.2.2.1 Si el tipo de atributo es numérico entonces el valor de ese atributo de la instancia sintética se calcula como sigue:
   a)$desviacion = min\{desviacion\_media, desviacion\_estandar\}$
   b)$Xrandom = valor\ aleatorio\ en\ [0, 1)$
   c)$Valor\_nuevo = valor\_original \pm (desviacion * Xrandom)$
   Las desviaciones estándar y media corresponden al grupo al que pertenece la instancia. En caso de ser un grupo con un único elemento se consideran las desviaciones globales.
   La operación de suma o resta ($\pm$) es aleatoria.
   1.2.2.2 Si el tipo de atributo es nominal entonces se verifica la proporción de valores presentes en el grupo. El porcentaje de cada valor en el grupo se tomará como una probabilidad y de acuerdo a ella se asignará el valor a la instancia sintética.

Tabla 4.4: Algoritmo GIS-GF.

Dado un conjunto de grupos, tal que n_{grupos} es el total de ellos y n_i es el tamaño del i-ésimo grupo, GIS-GF generará N instancias sintéticas distribuyéndolas de acuerdo al tamaño de cada clúster. Para el caso de atributos nominales el procedimiento es el mismo utilizado en GIS-G. Pero para crear un valor numérico el método es utilizar la menor de dos desviaciones, desviación estándar o desviación media, de los datos de ese grupo. Este valor se multiplica por un número aleatorio en $[0, 1]$ y es sumado o restado aleatoriamente al valor original para ser asignado al atributo de la instancia nueva sintética.

GIS-GF crea instancias sin considerar a los vecinos más cercanos (para el caso de valores numéricos) y cercanas a la instancia semilla. Un ejemplo se muestra en la figura 4.7, las instancias sintéticas están en el área circundante a la instancia original.

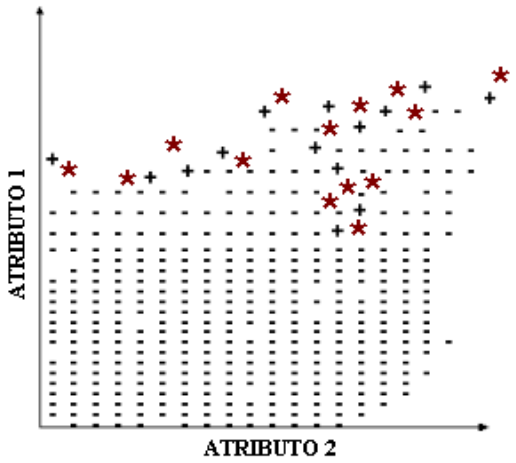

Figura 4.7: Ejemplo de generación de instancias con GIS-GF. Las instancias sintéticas están representadas por una signo asterisco.

4.4. Resumen

Se propusieron dos nuevas técnicas de generación inteligente de instancias llamadas GIS-G y GIS-GF. La idea principal en ambas es crear instancias sintéticas en áreas dominadas por la clase minoritaria y de manera local en lugar de global. Al considerar áreas para las instancias minoritarias no será necesario establecer un valor k para los vecinos más cercanos, ya que todas las instancias dentro de una misma región serán vecinas entre ellas. Para tratar el caso especial de los atributos nominales se propuso hacerlo probabilísticamente de acuerdo a la distribución de valores. Para el caso de atributos numéricos, GIS-G sigue el mismo método que SMOTE (interpolación) mientras que GIS-GF hace uso de la desviación estándar de los datos del grupo.

En el siguiente capítulo se describen las pruebas realizadas para comprobar la eficacia de GIS-G y GIS-GF con respecto a ROS, SMOTE y Borderline-SMOTE. Además se reportan los resultados obtenidos.

Capítulo 5

Experimentos y resultados

En este capítulo se describe la fase de pruebas de los métodos de sobre-muestreo propuestos en este trabajo (GIS-G y GIS-GF) comparados con los métodos de sobre-muestreo reportados en el estado del arte (ROS, SMOTE, Borderline-SMOTE1 y Borderline-SMOTE2). Los datos con que fueron probados los métodos se obtuvieron de veinte dominios artificiales y veintrés dominios del mundo real. A cada uno de ellos se les aplicó los seis diferentes métodos de sobre-muestreo utilizando validación cruzada. Los conjuntos sobre-muestreados se probaron con seis clasificadores diferentes proporcionados por la herramienta *Weka*. Las medidas de evaluación del desempeño utilizadas fueron las siguientes: recuerdo, precisión, Medida-F y AUC. Con el propósito de evitar que los resultados puedan atribuirse a la aleatoreidad introducida por los métodos de sobre-muestreo, el proceso completo de prueba se realizó diez veces para finalmente obtener el promedio de los resultados. Además se realizaron pruebas de *Análisis de Varianza* así como de *Significancia Estadística* para verificar la validez de los resultados obtenidos.

5.1. Bases de datos utilizadas

5.1.1. Bases de datos artificiales

Japkowicz y Stephen [14] encontraron que el grado de desbalance entre las clases, la complejidad del concepto y el tamaño del conjunto de entrenamiento son lo que más afecta a los clasificadores sensibles a este problema. En base a ello, se generaron diez bases de datos con dos atributos numéricos (bases de datos numéricas), cinco bases de datos con dos atributos nominales (bases de datos nominales) y cinco bases de datos con un atributo numérico y uno nominal (bases de datos combinadas). Todas las bases de datos tienen dos clases (minoritaria o postiva y mayoritaria o negativa).

Las diez bases de datos numéricas se generaron de la siguiente manera: cada una de las bases de datos pertenecen a un dominio bidimensional (una dimensión por cada atributo), cada atributo toma valores del rango [0, 1] y se asocia con una de dos clases, positiva (minoritaria) o negativa (mayoritaria). Hay cinco niveles de complejidad:

- Nivel 0: Sólo el rango de uno de los atributos se divide en dos para generar los valores correspondientes a cada clase, el otro atributo puede tomar cualquier valor en el rango completo. Esto es con la finalidad de tener problemas simples con sólo una región por cada clase. Una base de datos de este tipo se muestra en la figura 5.1 donde las instancias de la clase positiva (minoritaria) se ilustran con un signo + y las de la clase negativa (mayoritaria) con -. Para el atributo 1 sólo las instancias de la clase positiva pueden tomar valores en el rango [0, 0.5] y los de la clase negativa en el rango (0.5, 1], para el caso del atributo 2 no existe restricción.

- Niveles 1 a 4: c es el nivel de complejidad, el rango [0, 1] se divide en c+1 intervalos iguales. Entonces hay c^2 áreas en el dominio bidimensional para generar

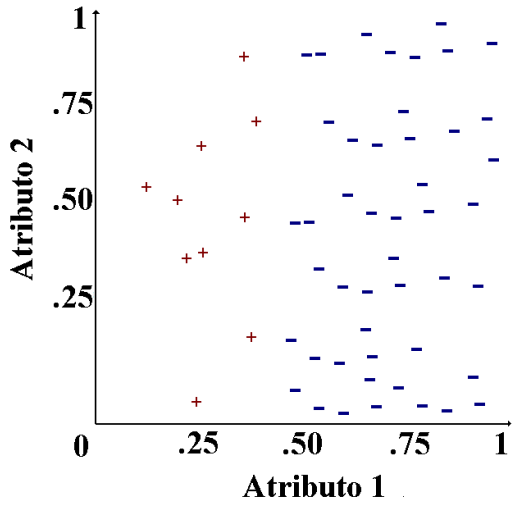

Figura 5.1: Base de datos de complejidad $c = 0$

instancias, las áreas contiguas tienen clases opuestas. Las instancias de cada área son generadas aleatoriamente y el grado de desbalance es variable. En la figura 5.2 se muestra una base de datos con complejidad $c = 1$.

Para generar las cinco bases de datos nominales los dos atributos fueron color (verde, azul, rojo y negro) y forma (rectángulo, cuadrado, círculo y óvalo). Cada atributo toma aleatoriamente un valor de los cuatro posibles.

Para bases de datos combinadas se utilizó para el primer atributo la misma técnica que para generar las bases de datos numéricas, para el segundo atributo la misma técnica que para bases de datos nominales.

Las características de todas las bases de datos artificiales se muestran en la tabla 5.1.

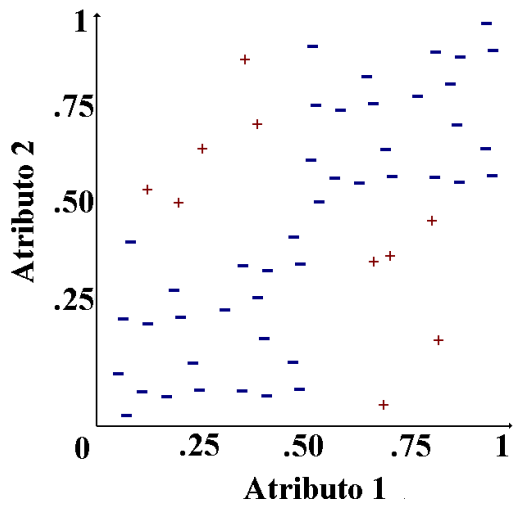

Figura 5.2: Base de datos de complejidad $c = 1$

Nombre	**Instancias positivas**	**Total instancias**	**% Positiva**	**Tipo de atributos**	**Comple-jidad**
Nominal01	10	110	9.09 %	Nominales	-
Nominal02	10	110	9.09 %	Nominales	-
Nominal03	20	120	16.67 %	Nominales	-
Nominal04	20	120	16.67 %	Nominales	-
Nominal05	50	300	16.67 %	Nominales	-
Numérica01	10	100	10 %	Numéricos	0
Numérica02	20	200	10 %	Numéricos	1
Numérica03	40	400	10 %	Numéricos	2
Numérica04	40	400	10 %	Numéricos	3
Numérica05	60	606	10 %	Numéricos	4
Numérica06	20	100	20 %	Numéricos	0
Numérica07	40	200	20 %	Numéricos	1
Numérica08	80	400	20 %	Numéricos	2
Numérica09	80	400	20 %	Numéricos	3
Numérica10	120	600	20 %	Numéricos	4
Combinada01	10	110	9.09 %	Combinados	0
Combinada02	10	110	9.09 %	Combinados	1
Combinada03	10	110	9.09 %	Combinados	2
Combinada04	10	110	9.09 %	Combinados	3
Combinada05	20	220	9.09 %	Combinados	4

Tabla 5.1: Bases de datos generadas artificialmente

5.1.2. Bases de datos reales

Se utilizaron veintitrés bases de datos tomadas del repositorio UCI [2], cada una de ellas de diferente tamaño, número y tipo de atributos y nivel de desbalance. Estas bases de datos originalmente tienen dos o más clases pero, para fines de estos experimentos, fueron convertidas a bases de datos con dos clases donde una es la clase positiva (minoritaria) y las demás clases se unen para formar la clase negativa (mayoritaria).

Algunas de las bases de datos presentaban incialmente el problema de datos faltantes. Dado que el propósito de esta tesis no es resolver tal problema, los valores desconocidos fueron reemplazados aleatoriamente por los diferentes valores del atributo correspondiente. Este proceso se realizó para la fase de preparación de los datos. Al aplicar los métodos de sobre-muestreo se asume que no hay datos faltantes.

Como se ha dicho, se aplican seis métodos de sobre-muestreo a diez capas de datos y eso se repite por diez veces. Para una sola base de datos se realizan en total seis cientas ejecuciones de sobre-muestreo. Además de que se llama a seis clasificadores diferentes en cada ejecución. Para probar los métodos de sobre-muestreo sobre una base de datos se realizan en total 600 operaciones de sobre muestreo (son seis métodos, diez capas, y el proceso completo se repite diez veces), sin considerar el tiempo para probar los seis clasificadores diferente. Entonces, en lo que concierne a recursos computacionales se consume mucho tiempo por lo que algunas bases de datos fueron reducidas en número de atributos y/o número de instancias. Las bases de datos reales utilizadas y sus características se muestran en la tabla 5.2.

5.2. Clasificadores

El problema de clases desbalanceadas puede afectar de diferente manera a cada clasificador, por ello las pruebas se realizaron con seis clasificadores de diferente tipo. Los parámetros utilizados en cada clasificador corresponden a los establecidos por default por *Weka*.

- *AdaBoost M1*: Utiliza árboles de decisión de un nivel (DecisionStump) y realiza diez iteraciones de boosting. El umbral de pesos se establece en 100.
- *Naive Bayes*: Utiliza precisión de 0.1 para atributos numéricos.
- *K-NN*: Considera los K=3 vecinos más cercanos.
- *C4.5*: No se utiliza poda y son mínimo dos instancias por hoja del árbol.
- *PART*: Utiliza C4.5 con un umbral de 0.25 para poda.
- *Backpropagation*: Tiene una tasa de aprendizaje 0.3 y momentum de 0.2. El número de capas ocultas se calcula como $a = (TotalClases{+}TotalAtributos)/2$ y el número de neuronas por capa es 4.

5.3. Análisis de Varianza

Para comprobar que los resultados obtenidos de los métodos de sobre-muestreo no son consecuencia del azar se realizó la prueba ANOVA con el método de sobre-muestreo como único factor. Los $a = 6$ diferentes niveles del factor (tratamientos) corresponden a los seis métodos de sobre-muestreo implementados. Son diez observaciones, las cuales son los resultados de las iteraciones que se realizan (recuerdo, precisión, Medida-F y AUC, cada una por separado). Con los valores de la tabla ANOVA

Nombre	Clase positiva	Instancias positivas	Total instancias	% Positiva	Num. de atributos	Tipo de atributos
Abalone1	4	100	600	16.66 %	9	Combinados
Balloons	T	35	76	46.05 %	5	Nominales
Breast Cancer	1	47	198	23.74 %	33	Numéricos
Bupa	1	145	345	42.03 %	7	Numéricos
Car1	good	70	700	10 %	7	Nominales
Car2	good	70	700	10 %	7	Nominales
Cardio-vascular	1	55	312	17.63 %	6	Numéricos
Cinco	V	20	106	18.87 %	3	Numéricos
Coil	autonum	36	184	19.57 %	18	Combinados
Cpu	hp	7	162	4.32 %	9	Numéricos
Diabetes	1	268	768	35 %	9	Numéricos
Ecoli	imU	35	332	10.54 %	8	Combinados
Escalon	V	19	100	19 %	3	Numéricos
Glass	3	17	214	7.94 %	10	Numéricos
Haberman	2	81	306	26.47 %	4	Numéricos
Hayes-Roth	3	30	132	22.72 %	4	Nominales
Heart	0	55	267	20.60 %	23	Nominales
Hepatitis	die	26	128	20.31 %	20	Combinados
Imayuscula	V	12	100	12 %	3	Numéricos
Post-operative patient	S	24	88	27.27 %	9	Combinados
Raro	V	24	82	29.26 %	3	Numéricos
Servo	B	50	161	31 %	5	Nominales
Wine	3	48	178	27 %	14	Numéricos

Tabla 5.2: Bases de datos de prueba tomadas del repositorio UCI.

que se creó se realizaron las sumas de cuadrados correspondientes y a partir de ello se obtuvo como conclusión si los resultados obtenidos son o no son resultado de la aleatoriedad introducida por los métodos de sobre-muestreo.

5.4. Significancia estadística

Todos los resultados de los métodos de sobre-muestreo son comparados entre sí para determinar si existe significancia estadística entre ellos. Por cada pareja comparada, con respecto al desempeño original de un clasificador, se realiza la prueba de significancia estadística de la siguiente manera:

Se debe tomar cada par de métodos de sobre-muestreo y probar sus medidas de desempeño (recuerdo, precisión, Medida-F o AUC) con respecto a la original (datos sin sobre-muestreo). La medida de desempeño original es σ. Se aplica cada método a una base de datos en diez iteraciones. Si el promedio de ellas para el primer método de sobre-muestreo es $x1'$ y para el segundo es $x2'$ entonces se aplica el siguiente procedimiento para verificar si existe significancia estadística. Se establece la hipótesis nula:

$$H_0 : \mu 1 - \mu 2 = \Delta_0 \tag{5.1}$$

significa que la diferencia entre las medias es igual a un parámetro establecido, en este caso $\Delta_0 = 0$. Se establece la hipótesis alternativa:

$$H_1 : \mu 1 - \mu 2 \neq \Delta_0 \tag{5.2}$$

Al calcular el estadístico de la prueba y estableciendo el criterio de rechazo como se explicó en el capítulo 2 (Marco Teórico) se determina si la diferencia en resultados obtenidos por los métodos de sobre-muestreo es o no significativa.

Los resultados generales de los experimentos realizados con los métodos de sobre-muestreo aparecen en los apéndices A al H, cada apéndice reporta una medida de desempeño sobre los seis clasificadores. El contenido de cada apéndice se lista a continuación:

- Apéndice A: Resultados del recuerdo sobre bases de datos artificiales.
- Apéndice B: Resultados de la precisión sobre bases de datos artificiales.
- Apéndice C: Resultados de la Medida-F sobre bases de datos artificiales.
- Apéndice D: Resultados del AUC sobre bases de datos artificiales.
- Apéndice E: Resultados del recuerdo sobre bases de datos reales.
- Apéndice F: Resultados de la precisión sobre bases de datos reales.
- Apéndice G: Resultados de la Medida-F sobre bases de datos reales.
- Apéndice H: Resultados del AUC sobre bases de datos reales.

La tabla 5.3 es un ejemplo de las tablas que se encuentran en los apéndices. En este caso la tabla reporta los resultados del recuerdo de la clase minoritaria que obtiene el clasificador *AdaBoost M1*. La primera columna indica el nombre de la base de datos a la que se aplicaron los métodos, los nombres en este ejemplo corresponden a las bases de datos generadas artificialmente. La segunda columna especifica el porcentaje de sobre-muestreo que se aplicó a la clase minoritaria, por ejemplo, la clase minoritaria de la base de datos "Combinada02" fue sobre-muestreada en un 900 %. La tercera columna, etiquetada como "original", se refiere al valor de la medida de desempeño (en este caso recuerdo) sobre la clase minoritaria que obtiene el clasificador sin aplicar sobre-muestreo. En el caso de la base de datos "Combinada02", el recuerdo de la clase minoritaria inicialmente es 0. Las últimas seis columnas indican el valor de la medida de desempeño utilizada, correspondiente a la clase minoritaria, que obtuvo

cada uno de los seis métodos de sobre-muestreo. Para el mismo ejemplo de la base de datos “Combinada02” los resultados del recuerdo para ROS, SMOTE, Borderline-SMOTE1 (BSM1), Borderline-SMOTE2(BSM2), GIS-G y GIS-GF son 0.640, 0.680, 0.250, 0.270, 0.730 y 0.740, respectivamente. Delante del nombre de la base de datos puede aparecer ** que indica que ANOVA mostró que los resultados para esa base de datos tienen significancia o sea que no pueden ser producto de la aleatoriedad introducida, la falta de estos símbolos en las bases de datos “Combinada04”, “Numérica01” y “Numérica06” indica que no pudo demostrarse que los resultados no son producto de la aleatoriedad. Por otro lado, un * delante de cada valor indica que la diferencia con GIS-G es estadísticamente significativa y un signo + significa que lo es con respecto a GIS-GF. Así, en la base de datos “Combinada02”, la diferencia de GIS-G y GIS-GF con los demás métodos es estadísticamente significativa pero la diferencia entre ellos no lo es.

Nombre	**%sob**	**Original**	**ROS**	**SMOTE**	**BSM1**	**BSM2**	**GIS-G**	**GIS-GF**
Nominal01**	900	0.000	0.490	0.440*+	0.030*+	0.030*+	0.500	0.500
Nominal02**	900	0.000	0.320	0.200*+	0.000*+	0.000*+	0.330	0.330
Nominal03**	400	0.000	0.530	0.430*+	0.035*+	0.035*+	0.515	0.515
Nominal04**	400	0.000	0.520	0.440*+	0.015*+	0.040*+	0.520	0.520
Nominal05**	400	0.000	0.622	0.482*+	0.010*+	0.010*+	0.624	0.624
Numérica01	800	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Numérica02**	800	0.000	0.185*+	0.695*+	0.605*+	0.675*+	0.690+	0.750*
Numérica03**	800	0.000	0.015*+	0.220*+	0.322*+	0.348*+	0.575+	0.588*
Numérica04**	800	0.000	0.038*+	0.057*+	0.180*+	0.173*+	0.248+	0.235*
Numérica05**	800	0.000	0.000*+	0.732*+	0.573*	0.517*	0.485+	0.525*
Numérica06	300	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Numérica07**	300	0.000	0.412*+	0.838*	0.810*	0.860*+	0.782+	0.815*
Numérica08**	300	0.000	0.094*+	0.150*+	0.408*+	0.381*+	0.571	0.580
Numérica09**	300	0.000	0.067*+	0.192*+	0.274*+	0.248*+	0.598	0.615
Numérica10**	300	0.000	0.002*+	0.182*+	0.257*+	0.267*+	0.614	0.634
Combinada01**	900	0.900	0.900+	0.900+	0.900+	0.900+	0.900+	1.000*
Combinada02**	900	0.000	0.640*+	0.680*+	0.250*+	0.270*+	0.730	0.740
Combinada03**	900	0.900	0.610	0.670	0.430	0.530	0.660	0.810
Combinada04	900	1.000	0.720	0.730	0.730	0.750	0.710	0.750
Combinada05**	900	0.760	0.645	0.570	0.505	0.560	0.615	0.625

Tabla 5.3: Ejemplo de una tabla de los apéndices. Reporta los resultados del recuerdo en bases de datos artificiales utilizando el clasificador *AdaBoost*.

En las siguientes secciones se explican los resultados obtenidos de acuerdo las medidas de desempeño y al clasificador utilizado.

5.5. Resultados por medida de desempeño

Considerando que son veinte bases de datos artificiales y se utilizaron seis clasificadores sobre cada una de ellas, se tienen ciento veinte resultados por cada medida de desempeño para probar la eficacia de los seis métodos de sobre-muestreo (cuatro del estado del arte y dos propuestos en esta tesis). Para el caso de las veintitrés bases de datos reales se obtienen 138 resultados.

5.5.1. Resultados del recuerdo

El recuerdo indica qué porcentaje de instancias de una clase dada fueron clasificadas correctamente. Esta medida es la de mayor interés en este trabajo, ya que el principal objetivo de los métodos de sobre-muestreo propuestos es mejorar la clasificación de las instancias minoritarias. Entonces, un recuerdo de GIS-G y GIS-GF que sea mayor a las obtenidas por los métodos ROS, SMOTE, BSM1 y BSM2 es un buen indicador de que los métodos propuestos son mejores que ellos.

Los apéndices A y E contienen los resultados del recuerdo de los seis clasificadores utilizados con las veinte bases de datos artificiales y reales respectivamente. Las tablas 5.4 (datos artificiales) y 5.5 (datos reales) muestran las veces que cada método de sobre-muestreo obtiene el resultado más alto sobre los demás métodos. En caso de existir un empate entre métodos se contabiliza como una victoria, por lo que al sumar las cantidades de la tabla puede suceder que el resultado sea mayor a ciento veinte.

Como se ve, en las bases de datos artificiales los métodos GIS-G y GIS-GF obtienen un número de victorias alto comparados con los demás métodos. En las bases de datos reales el método que obtiene más victorias es BSM1 seguido por GIS-GF y GIS-G. En principio esto puede tomarse como referencia para probar la eficacia de GIS-G y GIS-GF. Pero, analizando más detalladamente, esta forma de contabilizar

Método	Victorias
ROS	23
SMOTE	12
BSM1	6
BSM2	17
GIS-G	37
GIS-GF	79

Tabla 5.4: Número de veces que un método de sobre-muestreo resulta ganador globalmente en el recuerdo con bases de datos artificiales.

Método	Victorias
ROS	7
SMOTE	14
BSM1	8
BSM2	54
GIS-G	26
GIS-GF	41

Tabla 5.5: Número de veces que un método de sobre-muestreo resulta ganador globalmente en el recuerdo con bases de datos reales.

	ROS	SMOTE	BSM1	BSM2	GIS-G	GIS-GF
GIS-G	56/1	63/7	75/3	70/4	-	2/19
GIS-GF	67/1	80/6	84/1	81/1	19/2	-

Tabla 5.6: Comparación uno a uno de los resultados del recuerdo de los métodos de sobre-muestreo en bases de datos artificiales, se considera significancia estadística.

	ROS	SMOTE	BSM1	BSM2	GIS-G	GIS-GF
GIS-G	92/0	25/3	37/5	27/3	-	0/6
GIS-GF	95/1	30/0	40/4	27/4	6/0	-

Tabla 5.7: Comparación uno a uno de los resultados del Recuerdo de los métodos de sobre-muestreo en bases de datos reales, se considera significancia estadística.

no considera la significancia estadística en los resultados y se hace de manera global. Entonces también se incluirá la evaluación tomando cada método propuesto y haciendo una comparación uno a uno con los demás métodos y únicamente tomando en cuenta los casos donde la diferencia en los resultados haya sido significativa. Esta evaluación se resume en las tablas 5.6 y 5.7. En estas tablas, el valor que se encuentra en cada celda es el número de veces que el método de ese renglón es significativamente mejor que el método de esa columna. Por ejemplo, en la primera celda de la tabla 5.6 está la etiqueta 56/1 que significa que GIS-G fue mejor que ROS (considerando sólo cuando hay significancia estadística) en 56 ocasiones mientras que ROS fue significativamente mejor que GIS-G sólo en una ocasión.

De las tablas 5.6 y 5.7 se puede observar una gran ventaja de GIS-G y GIS-GF sobre cada uno de los demás métodos de sobre-muestreo. Aunque GIS-G en la mayoría de los casos no obtiene resultados estadísticamente significativos mejores con respecto a GIS-GF.

5.5.2. Resultados de la precisión

La precisión nos dice, de las instancias clasificadas como positivas cuántas en realidad lo son. Una precisión baja indica que hubo muchos falsos positivos (aunque todos los de la clase minoritaria hayan sido reconocidos correctamente). Es de esperarse que al mejorar la clasificación de la clase minoritaria con sobre-muestreo se afecte la clasificación de la clase mayoritaria. Esto se debe a que, como se mencionó en el capítulo 1, el sobre-muestrear la clase minoritaria de un conjunto de datos es equivalente a modificar los costos de clasificación favoreciendo a dicha clase. Esto provoca que se clasifiquen instancias de la clase mayoritaria como si fueran de la clase minoritaria, esto es válido dado que la clase de mayor interés es la clase positiva. Es preferible tener una mejor clasificación de instancias minoritarias aunque el costo sea sacrificar parte de las instancias mayoritarias. Sin embargo, lo ideal sería afectar en lo más mínimo a la clase mayoritaria.

En los apéndices B y F se encuentran los resultados de la precisión obtenidos de los seis clasificadores utilizados sobre las veinte bases de datos artificiales y las veintrés bases de datos reales, respectivamente. Vea las tablas 5.8 y 5.9 donde aparecen estos resultados. Al contabilizar las veces que cada método de sobre-muestreo obtuvo el valor más alto, podría concluirse erróneamente que SMOTE obtiene los mejores resultados en cuanto a la precisión. En el caso de bases de datos artificiales SMOTE obtiene el mejor resultado en 35 ocasiones y con bases de datos reales en 36, posiblemente empatando algunas veces con otros métodos, pero cuando es mejor, no es significativamente mejor que los métodos propuestos. Por otro lado, puede suceder que los métodos propuestos no sean los mejores para cierta base de datos, pero son significativamente mejores respecto a SMOTE. Entonces, realizando una comparación de GIS-G y GIS-GF contra todos los demás y contabilizando únicamente aquellos casos donde la diferencia fue estadísticamente significativa, obtenemos las tablas 5.10 y 5.11.

Método	Victorias
ROS	23
SMOTE	35
BSM1	18
BSM2	10
GIS-G	15
GIS-GF	15

Tabla 5.8: Número de veces que un método de sobre-muestreo resulta ganador globalmente en Precisión con bases de datos artificiales.

Método	Victorias
ROS	36
SMOTE	20
BSM1	16
BSM2	1
GIS-G	14
GIS-GF	13

Tabla 5.9: Número de veces que un método de sobre-muestreo resulta ganador globalmente en la precisión con bases de datos reales.

	ROS	SMOTE	BSM1	BSM2	GIS-G	GIS-GF
GIS-G	24/2	22/3	30/10	33/9	-	6/5
GIS-GF	26/3	25/3	35/10	35/9	5/6	-

Tabla 5.10: Comparación uno a uno de los resultados de la precisión de los métodos de sobre-muestreo en bases de datos artificiales, se considera significancia estadística.

	ROS	SMOTE	BSM1	BSM2	GIS-G	GIS-GF
GIS-G	26/14	13/0	29/12	38/1	-	5/4
GIS-GF	24/21	10/8	24/19	37/5	4/5	-

Tabla 5.11: Comparación uno a uno de los resultados de la precisión de los métodos de sobre-muestreo en bases de datos reales, se considera significancia estadística.

En las tablas 5.10 y 5.11 un valor del tipo x/y en la celda (a, b) significa que el método del renglón a fue mejor que el método de la columna b en x ocasiones, y que el método b fue mejor que el método a en y ocasiones. Por ejemplo, según la tabla 5.10, GIS-GF fue mejor que BSM1 en 35 ocasiones mientras que BSM1 fue mejor que GIS-GF en 10 ocasiones, así el valor reportado en la tabla es 35/10. Como puede observarse, GIS-G y GIS-GF tienen la mayor precisión comparados con los demás métodos en la mayoría de las ocasiones y entre ellos la diferencia es muy poca.

5.5.3. Resultados de la Medida-F

En los apéndice C (bases de datos artificiales) y G (bases de datos reales) se reportan los resultados de la Medida-F obtenidos en la fase de experimentos con los seis clasificadores. La Medida-F es una medida de compromiso entre recuerdo y precisión, entonces una valor alto en la Medida-F indica que tanto recuerdo como precisión tienen valores altos.

Del mismo modo que para recuerdo y precisión, contabilizamos las veces que un método obtiene los valores de la Medida-F más altos comparados globalmente.

Método	Victorias
ROS	26
SMOTE	35
BSM1	23
BSM2	10
GIS-G	13
GIS-GF	34

Tabla 5.12: Número de veces que un método de sobre-muestreo resulta ganador globalmente en Medida-F con bases de datos artificiales.

Método	Victorias
ROS	12
SMOTE	28
BSM1	23
BSM2	13
GIS-G	28
GIS-GF	19

Tabla 5.13: Número de veces que un método de sobre-muestreo resulta ganador globalmente en la Medida-F con bases de datos reales.

Vea las tablas 5.12 y 5.13. Observe que, en el caso de datos artificiales, SMOTE gana treinta y cinco veces y que GIS-GF en treinta y cuatro. Con datos reales SMOTE y GIS-G obtienen 28 victorias.

Ahora veamos la comparación por pares de métodos que se muestra en las tablas 5.14 y 5.15. Aquí se puede ver nuevamente que los métodos propuestos obtienen mejores resultados (estadísticamente significativos) en muchas más ocasiones que los métodos ROS, SMOTE, BSM1 y BSM1 y que entre ellos no hay mucha diferencia, aunque GIS-GF muestra una pequeña ventaja sobre GIS-G.

	ROS	SMOTE	BSM1	BSM2	GIS-G	GIS-GF
GIS-G	48/2	35/10	58/2	59/1	-	9/7
GIS-GF	48/4	40/10	57/4	56/3	7/9	-

Tabla 5.14: Comparación uno a uno de los resultados de la Medida-F de los métodos de sobre-muestreo en bases de datos artificiales, se considera significancia estadística.

	ROS	SMOTE	BSM1	BSM2	GIS-G	GIS-GF
GIS-G	47/0	8/0	17/3	21/3	-	1/0
GIS-GF	47/2	4/1	14/5	22/3	0/1	-

Tabla 5.15: Comparación uno a uno de los resultados de la Medida-F de los métodos de sobre-muestreo en bases de datos reales, se considera significancia estadística.

5.5.4. Resultados del AUC

El área bajo la curva ROC es útil para conocer el rendimiento global de una prueba. Para propósitos de esta tesis estamos más interesados en el desempeño de los métodos de sobre-muestreo sobre la clase positiva, aunque sin hacer a un lado el desempeño global de los clasificadores. Entonces, analizando los resultados en los apéndices D (bases de datos artificiales) y H (bases de datos reales), con bases de datos artificiales GIS-GF obtiene el AUC más alta en más ocasiones, seguido por BSM1 (tabla 5.16). Con bases de datos reales SMOTE es quien tiene mejores resulados en más ocasiones (tabla 5.17). Pero una vez más se requiere mostrar la eficacia de GIS-G y GIS-GF sobre los demás métodos y con la presencia de significancia estadística en los resultados. Realizando ese análisis los valores obtenidos se reportan en las tablas 5.18 y 5.19.

De las tablas 5.18 y 5.19 se ve que son pocas las ocasiones que existe siginificancia estadística con respecto a los otros médodos pero en las ocasiones que la hay GIS-G y GIS-GF son los mejores. Podemos decir que GIS-G y GIS-GF tienen un desempeño en cuanto a AUC al menos tan bueno como los otros métodos, incluso mejor.

Método	Victorias
ROS	19
SMOTE	11
BSM1	31
BSM2	20
GIS-G	21
GIS-GF	33

Tabla 5.16: Número de veces que un método de sobre-muestreo resulta ganador globalmente en la medida AUC con bases de datos artificiales.

Método	Victorias
ROS	29
SMOTE	33
BSM1	11
BSM2	8
GIS-G	27
GIS-GF	15

Tabla 5.17: Número de veces que un método de sobre-muestreo resulta ganador globalmente en la medida AUC con bases de datos reales.

	ROS	SMOTE	BSM1	BSM2	GIS-G	GIS-GF
GIS-G	12/0	0/0	9/0	9/0	-	0/0
GIS-GF	12/0	0/0	9/0	9/0	0/0	-

Tabla 5.18: Comparación uno a uno de los resultados del AUC de los métodos de sobre-muestreo en bases de datos artificiales, se considera significancia estadística.

	ROS	SMOTE	BSM1	BSM2	GIS-G	GIS-GF
GIS-G	16/0	4/0	2/0	3/0	-	0/0
GIS-GF	16/0	4/0	2/0	3/0	0/0	-

Tabla 5.19: Comparación uno a uno de los resultados del AUC de los métodos de sobre-muestreo en bases de datos reales, se considera significancia estadística.

En el trabajo desarrollado por Jason Van Hulse, Taghi M. Khoshgoftaar y Amri Napolitano [28] se concluye que ROS, en general, muestra mejores resultados que los métodos SMOTE y Borderline-SMOTE. En esta sección, los resultados de todas las medidas de desempeño muestran que los métodos GIS-G y GIS-GF son significativamente mejores que ROS.

5.6. Resultados por clasificador

En la sección anterior se mostró la eficacia de GIS-G y GIS-GF sobre los métodos de sobre-muestreo presentados en el estado del arte. En esta sección se reportan los resultados que obtienen los dos métodos propuestos considerando el clasificador utilizado. De esta manera se vericará el comportamiento de GIS-G y GIS-GF con cada clasificador.

5.6.1. GIS-G

La tabla 5.20, obtenida de los apéndices, presenta el total de veces que GIS-G obtuvo el mejor resultado de cada medida de desempeño de acuerdo al clasificador utilizado. Utilizando C4.5 GIS-G obtiene el mejor recuerdo (empata con *Backpropagation*), precisión, Medida-F y AUC (empata con K-NN y *Backpropagation*). De ahí se puede concluir que para obtener los mejores resultados cuando se utilice el clasificador C4.5 el mejor método de sobre-muestreo es GIS-G.

	AdaBoost M1	Naive Bayes	K-NN	C4.5	PART	Backpropagation
Recuerdo	9	9	11	12	10	12
Precisión	5	5	5	6	3	6
Medida-F	5	8	6	10	6	7
AUC	4	4	11	11	7	11

Tabla 5.20: Número de veces GIS-G resulta ganador con cada clasificador.

	AdaBoost M1	Naive Bayes	K-NN	C4.5	PART	Backpropagation
Recuerdo	18	19	21	21	19	22
Precisión	4	7	6	4	5	4
Medida-F	9	10	7	10	8	10
AUC	5	5	9	11	8	10

Tabla 5.21: Número de veces GIS-GF resulta ganador con cada clasificador.

5.6.2. GIS-GF

La tabla 5.21, obtenida de los apéndices, presenta el total de veces que GIS-GF obtuvo el mejor resultados de cada medida de desempeño de acuerdo al clasificador utilizado. En este caso, los resultados son más variables que con GIS-G. Si estamos interesados en mejorar el recuerdo entonces con *Backpropagation* es con el que obtiene los mejores resultados. Para precisión GIS-GF con *Naive Bayes* resulta la mejor combinación. Para Medida-F, que es un compromiso entre los dos anteriores, se recomienda utilizar GIS-GF con *Naive Bayes*, C4.5 o *Backpropagation*. Por último, para AUC es mejor GIS-GF con C4.5. Por otro lado, los clasificadores que obtienen los resultados más bajos utilizando GIS-GF son: para recuerdo es *AdaBoost M1*, para Precisión son *AdaBoost M1*, C4.5 y *Backpropagation*, para Medida-F es K-NN y para AUC son *AdaBoost M1* y *Naive Bayes*. Con PART no se obtienen ni los mejores ni los peores resultados. De lo anterior podemos decir que cuando se utilicen los clasificadores *AdaBoost M1* y PART se recomienda no utilizar GIS-GF.

5.7. Resultados por tiempo

Conocer el tiempo de ejecución de un algoritmo sirve como una referencia más para probar su eficacia. El objetivo de los algoritmos de sobre-muestreo es mejorar la clasificación de instancias minoritarias. Sin embargo, debe considerarse el tiempo que se requiere para llevar a cabo la generación de instancias, ya que éste no debe ser excesivo.

Para calcular el tiempo que tarda cada algoritmo de sobre-muestreo (cuatro del estado del arte y dos propuestos) se realizó la ejecución de cada uno de ellos. Debido a la aleatoriedad utilizada en los métodos, se repitió la ejecución diez veces y se promediaron los resultados. Estas ejecuciones se llevaron a cabo en una computadora con las siguientes características:

- Procesador Pentium 4
- CPU a 3.20 GH
- 1 GB de RAM

El tiempo, contabilizado en milisegundos, sólo incluye la generación de instancias sintéticas, no se contabiliza el tiempo que toma adquirir los datos de un archivo .arff ni mostrar o guardar los resultados.

El tiempo que tardó cada método con cada base de datos se muestra en la tabla 5.22. Como puede observarse, Borderline-SMOTE2 es el método que más tiempo requiere para su ejecución, mientras que ROS y SMOTE son a quienes menos tiempo lleva sobre-muestrear un conjunto de datos. El tiempo utilizado por GIS-G y GIS-GF se mantiene entre el tiempo requerido por Borderline-SMOTE1 y Borderline-SMOTE2.

Nombre	**ROS**	**SMOTE**	**BSM1**	**BSM2**	**GIS-G**	**GIS-GF**
Nominal01	31	34	65,496	100,642	67,394	85,507
Nominal02	34	35	82,533	113,732	74,774	75,841
Nominal03	35	36	90,036	124,071	81,572	82,736
Nominal04	36	36	90,035	124,068	81,572	82,736
Nominal05	90	90	225,088	310,170	203,930	206,840
Numérica01	39	40	55,083	80,450	58,091	72,903
Numérica02	78	78	110,172	160,917	116,178	145,800
Numérica03	155	157	220342	321836	232355	291601
Numérica04	159	164	220,347	321,829	232,350	291,600
Numérica05	243	245	333,826	487,571	352,010	441,774
Numérica06	39	15	55,085	80,457	58,100	72,899
Numérica07	79	32	110,173	160,915	116,176	145,802
Numérica08	157	158	220344	321832	232351	291605
Numérica09	159	160	220,347	321,829	232,350	291,600
Numérica10	243	245	331,071	483,548	349,106	438,129
Combinado01	39	39	73,411	114,687	77,341	78,055
Combinado02	46	45	68,505	105,072	72,909	68,409
Combinado03	39	40	70,580	109,279	67,733	71,497
Combinado04	39	39	71,588	109,661	72,662	71,363
Combinado05	85	86	139,085	214,351	140,642	139,906

Tabla 5.22: Tiempo promedio en milisegundos que tarda un método de sobre-muestreo.

5.8. Resumen

En este capítulo presentaron las pruebas realizadas a los métodos de sobremuestreo GIS-G y GIS-GF comparados con ROS, SMOTE, Borderline-SMOTE1 y Borderline-SMOTE2. Los resultados mostraron que los dos métodos propuestos en general obtienen mejores resultados (mejoran la clasificación de instancias minoritarias) que los reportados en el estado del arte. Estos resultados están respaldados por análisis de varianza y de significancia estadística.

El siguiente capítulo contiene las conclusiones que se derivan de estos experimentos y el trabajo futuro desprendido de esta tesis.

Capítulo 6

Conclusiones y trabajo futuro

En este capítulo se presentan las conclusiones que se desprenden del trabajo desarrollado y de los resultados obtenidos. También se propone el trabajo futuro a realizar.

6.1. Conclusiones

Se propusieron dos métodos nuevos de sobre-muestreo para tratar el problema de clases desbalanceadas GIS-G y GIS-GF. Éstos consideran la distribución de las instancias y generan grupos de ellas para así crear las instancias sintéticas dentro de cada uno de estos grupos. Para el caso de atributos numéricos GIS-G utiliza interpolación de valores de dos instancias para asignar el valor a la nueva instancia y GIS-GF utiliza sólo una instancia como semilla y considera la desviación estándar de los valores del grupo correspondiente. En cuanto a atributos nominales, ambos métodos consideran la distribución de valores nominales dentro de cada grupo, el valor asignado a la nueva instancia será acorde a esa distribución y de manera probabilista. Con la formación

de grupos se definen mejor las áreas influenciadas por las clase positiva y se generan los valores nominales de una forma más equitativa de acuerdo a la distribución de los valores dentro de un grupo. Las ventajas obtenidas son:

- Las instancias sintéticas creadas no afectan las áreas dominadas por la clase negativa ya que se generan dentro del grupo que define el área de la clase minoritaria. Esta es una forma novedosa de generación de instancias.
- No es necesario definir el valor de k para la selección de vecinos más cercanos ya que se utilizan como vecinas las instancias que pertenecen al mismo grupo.
- En el caso de GIS-GF, para valores numéricos sólo se requiere de una instancia como semilla.
- La generación de valores nominales para las instancias sintéticas se hace de una forma equitativa deacuerdo a la distribución de valores y no únicamente repitiéndolos.
- No es necesario que las fronteras entre clases estén bien definidas como en el caso de los métodos Borderline-SMOTE.

Ambos métodos lograron incrementar el recuerdo (tasa de verdaderos positivos), Medida-F y mantener el área bajo la curva ROC (AUC) de la clase minoritaria. Pruebas de significancia estadística y análisis de varianza avalan los resultados. Estos métodos mostraron ser mejores que los métodos reportados en el estado del arte (ROS, SMOTE, BSM1 y BSM2) en la mayoría de los casos.

Las pruebas muestran que en varias ocasiones ROS presenta un desempeño mejor sobre los demás métodos de sobre-muestreo que se presentan en el estado del arte, esto es acorde con los resultados obtenidos del trabajo de Jason Van Hulse, Taghi M. Khoshgoftaar y Amri Napolitano [28]. Sin embargo, GIS-G y GIS-GF logran superarlo.

Cuando SMOTE genera instancias sintéticas a partir de una instancia de la clase positiva con uno de sus vecinos más cercanos, las instancias sintéticas se generan en regiones donde predominan instancias de la clase negativa. Esto ocasiona que sea más difícil para el clasificador encontrar un modelo que pueda clasificar esas instancias. GIS-G y GIS-GF no presentan este problema.

BSM1 y BSM2 muestran buenos resultados cuando los bordes de las clases están bien delimitados, pero para que esto suceda las instancias de los bordes de cada clase deben estar más cercanas entre ellas que con las instancias de su misma clase, de no ser así no son consideradas en el borde y no se lleva a cabo el sobre-muestreo de ellas. GIS-G y GIS-GF encuentran grupos sin importar la distancia que exista entre los bordes de las clases.

Cuando se utilice el clasificador C4.5 es recomendable utilizar el método de sobre-muestreo GIS-G. Mientras que con AdaBoost ni PART no se recomienda utilizar GIS-GF. Con los otros clasificadores no hubo diferencia marcada en el número de veces que obtuvieron los mejores resultados.

6.2. Trabajo futuro

Existen varios trabajos a desarrollar como consecuencia de esta tesis que se proponen como trabajo futuro y se presentan en esta sección.

Un problema frecuente en las tareas de clasificación es el ruido en bases de datos. Sin embargo, no se conoce el comportamiento de los diferentes métodos de sobre-muestreo ante este problema. Por eso se requiere probar el efecto del ruido en los métodos de sobre-muestreo presentados en esta tesis.

También es necesario experimentar con otros criterios de fusión de grupos con el objetivo de definir de manera más precisa las regiones que comprende la clase positiva. En esta tesis, para fusionar dos grupos, se verifica si hay más instancias de la clase minoritaria que de la mayoritaria. Pero podría darse el caso de que exista una cantidad más o menos igual de ambas instancias. Pueden buscarse opciones para que el criterio de fusión no sea únicamente verificando si hay más instancias de una clase que de otra. Hay que probar con diferentes proporciones, por ejemplo *si más de 2/3 de las instancias pertenecen a la clase minoritaria entonces se fusionan los grupos*.

Como puede observarse en algunas tablas de resultados, hay algunas ocasiones en que el sobre-muestreo (sin importar el método utilizado) resulta perjudicial a la clasificación. Este es un asunto que debe ser investigado con mayor profundidad y detalle.

Por otro lado, para el caso de re-muestreo existe otra vertiente que es el sub-muestreo. En base a GIS-G y GIS-GF pueden desarrollarse métodos de sub-muestreo que aprovechen la formación de grupos y en base a ello decidir cuáles instancias deben ser removidas del conjunto de entrenamiento. Una opción sería remover las instancias que queden en el centro del grupo y conservas las demás que podrían considerarse como bordes.

Combinar métodos de re-muestreo, sobre-muestreando la clase positiva y sub-muestreando la clase negativa, es otra alternativa que parece prometedora para lidiar con el problema de clases desbalanceadas. Entonces es necesario probar y analizar las combinaciones posibles de estos métodos de tal forma que pueda encontrarse aquélla que en general obtenga el mejor desempeño de los clasificadores.

Bibliografía

[1] G. E. A. P. A. Batista, R. C. Prati yM. C. Monard. A study of the behavior of several Methods for balancing machine learning training data. SIGKDD Explorations, 6(1):20-29, 2004.

[2] C. Blake, y C. Merz. UCI Repository of Machine Learning Databases http://www.ics.uci.edu/mlearn/MLRepository.html. Department of Information and Computer Sciences, University of California, Irvine, 1998.

[3] D. R. Carvalho y A. A. Freitas. A genetic algorithm for discovering small-disjunct rules in data mining. Applied Soft Computing pp. 75-88, 2002.

[4] Nitesh V. Chawla, K. W. Bowyer, L. O. Hall y W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. JAIR 16, pp. 321-357, 2002.

[5] Nitesh V. Chawla, Nathalie Japkowicz y Aleksander Kolcz. Editorial: Special Issue on Learning from Imbalanced Data Sets. SIGKDD Explorations 6 (1), pp. 1-6, 2004.

[6] C.H. Chou, B.H. Kuo y F. Chang. The Generalized Condensed Nearest Neighbor Rule as a Data Reduction Method. 18th International Conference on Pattern Recognition (ICPR06). Hong Kong, China, pp. 556-559, 2006.

[7] P. Domingos y M. Pazzani. MetaCost: A General Method for Making Classifiers Cost-Sensitive. In Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99), 1999.

[8] Chris Drummond y Robert C. Holte. Cost Curves: An Improved Method for Visualizing Classifier Performance. Machine Learning, 2006.

[9] Duda, O. Richard , Stork, G. David, Peter E. Hart. Pattern Classification and Scene Analysis: Pattern Classification ,Wiley. 2000.

[10] K. J. Ezzawa, M. Singh y S. W. Norton. Learning Goal Oriented Bayesian Networks for Telecommunications Managment. Proceedings of the International Conference on Machine Learning, ICML'96 , Bari, Italy, Morgan Kafmann, pp. 139-147, 1996.

[11] T. Fawcett y F. Provost. Combining Data Mining and Machine Learning for Effective User Profile. Proceeding of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland OR, AAAI Press, pp. 8-13, 1996.

[12] Yoav Freund y Robert E. Schapire. A decision-theoretic generalization of online learning and an application to boosting. En Proc. 2nd European Conference on Computational Learning Theory, pp. 23-37, 1995.

[13] H. Han, W. Y. Wang y B. H. Mao. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In International Conference on Intelligent Computing (ICIC'05). Lecture Notes in Computer Science 3644. Springer-Verlag, pp. 878-887, 2005.

[14] N. Japkowicz y S. Stephen. The Class Imbalance Problem: A Systematic Study. Intelligent Data Analysis, pp. 429-450, 2002.

[15] M. Kubat y S. Matwin. Addressing the Curse of Imbalanced Training Sets: One-sided Selection. In ICML'97, pp. 179-186, 1997.

[16] J Laurikkala. Improving Identification of Difficult Small Classes by Balancing Class Distribution. Tech. Rep. A-2001-2, University of Tampere, 2001.

[17] Maloof y Marcus A. Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown. Workshop on Learning from Imbalanced Data Sets II at the International Conference on Machine Learning, 2003.

[18] Luis Mena y Jesús A. González. Machine Learning for Imbalanced Datasets: Application in Medical Diagnostic. En Proc. of the Nineteenth International Florida Artificial Intelligence Research Society, AAAI Press, 2006.

[19] R. S. Michalski e I. Bratko, M. Kubat. Machine Learning and Data Mining, Methods and applications, Ed. Wiley, pp. 389-403, 1998.

[20] Douglas C. Montgomery y George C. Runger. Applied Statistics and Probability for engineers. Third edition. Wiley. 2003.

[21] R. Quinlan. C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.

[22] P. Riddle, R. Segal y O. Etzioni. Representation design and bruteforce induction in a Boeing manufacturinf design. Applied Artificial Intelligence, pp. 125-147, 1994.

[23] R. E. Schapire. The Boosting Approach to Machine Learning: An Overview. MSRI Workshop on Nonlinear Estimation and Classification, pp. 1-23, 2002.

[24] A.V. Sousa, A.M. Mendonca y A. Campilho, The Class Imbalance Problem in TLC Image Classification, Image Analysis and Recognition, LNCS 4142, pp. 513-523, 2006.

[25] E. Stamatatos. Text Sampling and Re-sampling for Imbalanced Author Identification Cases, In Proc. of the 17th European Conference on Artificial Intelligence (ECAI'06), 2006.

[26] K. M. Ting. The problem of small disjuncts: its remedy in decision trees. En Proceedings of the Tenth Canadian Conference on Artificial Intelligence, pp. 91-97, 1994.

[27] I. Tomek. Two Modifications of CNN. IEEE Transactions on Systems Man and Communications SMC-6, pp. 769-772, 1976.

[28] Jason Van Hulse, Taghi M. Khoshgoftaar y Amri Napolitano. Experimental perspectives on learning from imbalanced data. ICML pp. 935-942, 2007.

[29] F. Vilariño, P. Spyridonos, J. Vitria y P. Radeva, Experiments with SVM and Stratified sampling with an Imbalanced Problem: Detecttion of Intestinal Contractions, S. Singh et al. (Eds.): ICAPR 2005, LNCS 3687, Springer-Verlag, (ISI 0, 402), pp. 783-791, 2005.

[30] D. L. Wilson. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. IEEE Transactions on Systems, Man, and Communications, pp. 408-421, 1972.

[31] D.R. Wilson y T.R. Martínez, Improved Heterogeneous Distance Functions, Journal of Artificial Intelligence Research, 6, pp. 1-34, 1997.

[32] Ian H. Witten y Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Second edition, Morgan Kaufmann. pp. 41-52, 1999.

[33] B. Zadrozny y C. Elkan. Learining and Making Decisions When Costs and Probabilities are Both Unknown. Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining, 2001.

APÉNDICE A

Resultados del recuerdo sobre bases de datos artificiales

Este apéndice contiene las tablas de resultados del recuerdo de los seis métodos de sobre-muestreo (ROS, SMOTE, BSM1, BSM2, GIS-G y GIS-GF) con las veinte bases de datos generadas artificialmente. Son seis tablas, cada una de ellas corresponde a cada clasificador utilizado: *AdaBoost M1*, *Naive Bayes*, K-NN, C4.5, PART y *Backpropagation*.

Una marca ** delante del nombre de la base de datos indica que ANOVA mostró que los resultados en general son significativos, no se deben a la aleatoriedad introducida por los métodos de sobre-muestreo. Un signo * delante de cada valor del recuerdo indica que la diferencia de ese valor con el obtenido por GIS-G es significativa y un signo + que lo es con GIS-GF.

Nombre	**%sob**	**Original**	**ROS**	**SMOTE**	**BSM1**	**BSM2**	**GIS-C**	**GIS-CJ**
Nominal01**	900	0.000	0.490	0.440*+	0.030*+	0.030*+	0.500	0.500
Nominal02**	900	0.000	0.320	0.200*+	0.000*+	0.000*+	0.330	0.330
Nominal03**	400	0.000	0.530	0.430*+	0.035*+	0.035*+	0.515	0.515
Nominal04**	400	0.000	0.520	0.440*+	0.015*+	0.040*+	0.520	0.520
Nominal05**	400	0.000	0.622	0.482*+	0.010*+	0.010*+	0.624	0.624
Numérica01	800	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Numérica02**	800	0.000	0.185*+	0.695+	0.605*+	0.675*+	0.690+	0.750*
Numérica03**	800	0.000	0.015*+	0.220*+	0.322*+	0.348*+	0.575+	0.588*
Numérica04**	800	0.000	0.038*+	0.057*+	0.180*+	0.173*+	0.248+	0.235*
Numérica05**	800	0.000	0.000*+	0.732*+	0.573*	0.517*	0.485+	0.525*
Numérica06	300	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Numérica07**	300	0.000	0.412*+	0.838*	0.810*	0.860*+	0.782+	0.815*
Numérica08**	300	0.000	0.094*+	0.150*+	0.408*+	0.381*+	0.571	0.580
Numérica09**	300	0.000	0.067*+	0.192*+	0.274*+	0.248*+	0.598	0.615
Numérica10**	300	0.000	0.002*+	0.182*+	0.257*+	0.267*+	0.614	0.634
Combinada01**	900	0.900	0.900+	0.900+	0.900+	0.900+	0.900+	1.000*
Combinada02**	900	0.000	0.640*+	0.680*+	0.250*+	0.270*+	0.730	0.740
Combinada03**	900	0.900	0.610	0.670	0.430	0.530	0.660	0.810
Combinada04	900	1.000	0.720	0.730	0.730	0.750	0.710	0.750
Combinada05**	900	0.760	0.645	0.570	0.505	0.560	0.615	0.625

Tabla 1: Resultados del recuerdo con BD artificiales y clasificador AdaBoostM1.

Nombre	**%sob**	**Original**	**ROS**	**SMOTE**	**BSM1**	**BSM2**	**GIS-C**	**GIS-CJ**
Nominal01**	900	0.000	0.490	0.440*+	0.030*+	0.030*+	0.500	0.500
Nominal02**	900	0.000	0.320	0.200*+	0.000*+	0.000*+	0.330	0.330
Nominal03**	400	0.000	0.530	0.430*+	0.035*+	0.035*+	0.515	0.515
Nominal04**	400	0.000	0.520	0.440*+	0.015*+	0.040*+	0.520	0.520
Nominal05**	400	0.000	0.622	0.482*+	0.010*+	0.010*+	0.624	0.624
Numérica01	800	0.800	1.000	1.000	1.000	1.000	1.000	1.000
Numérica02**	800	0.000	0.185*+	0.695+	0.605*+	0.675+	0.690+	0.750*
Numérica03**	800	0.000	0.015*+	0.220*+	0.322*+	0.348*+	0.575	0.555
Numérica04**	800	0.000	0.038*+	0.057*+	0.180*+	0.173*+	0.248	0.235
Numérica05**	800	0.000	0.000*+	0.732*+	0.573	0.517	0.485+	0.525*
Numérica06	300	0.900	1.000	1.000	1.000	1.000	1.000	1.000
Numérica07**	300	0.000	0.412*+	0.838*	0.810	0.860*	0.782	0.815
Numérica08**	300	0.000	0.094*+	0.150*+	0.408*+	0.381*+	0.571	0.580
Numérica09**	300	0.000	0.067*+	0.192*+	0.274*+	0.248*+	0.598	0.615
Numérica10**	300	0.000	0.002*+	0.182*+	0.257*+	0.267*+	0.614	0.634
Combinada01**	900	0.700	0.900+	0.900+	0.900+	0.910+	0.900+	1.000*
Combinada02**	900	0.000	0.640*+	0.680*+	0.250*+	0.270*+	0.730	0.740
Combinada03**	900	0.000	0.610*+	0.670+	0.430*+	0.530*+	0.660+	0.810*
Combinada04**	900	0.000	0.720+	0.730+	0.730+	0.750*	0.710+	0.750*
Combinada05**	900	0.000	0.645*+	0.570*+	0.505*+	0.560*+	0.615+	0.625*

Tabla 2: Resultados del recuerdo con BD artificiales y clasificador *Naive Bayes*.

Nombre	**%sob**	**Original**	**ROS**	**SMOTE**	**BSM1**	**BSM2**	**GIS-C**	**GIS-CJ**
Nominal01**	900	0.000	0.490	0.440	0.030*+	0.030*+	0.500	0.500
Nominal02**	900	0.000	0.320	0.200*+	0.000*+	0.000*+	0.330	0.330
Nominal03**	400	0.000	0.530	0.430*+	0.035*+	0.035*+	0.515	0.515
Nominal04**	400	0.000	0.520	0.440*+	0.015*+	0.040*+	0.520	0.520
Nominal05**	400	0.002	0.622	0.482*+	0.010*+	0.010*+	0.624	0.624
Numérica01	800	0.600	1.000	1.000	1.000	1.000	1.000	1.000
Numérica02**	800	0.715	0.185*+	0.695	0.605	0.675	0.690	0.750
Numérica03**	800	0.458	0.015*+	0.220*+	0.322*+	0.348*+	0.575	0.588
Numérica04**	800	0.352	0.038*+	0.057*+	0.180	0.173	0.248	0.235
Numérica05**	800	0.227	0.000*+	0.732*+	0.573	0.517	0.485	0.525
Numérica06	300	0.905	1.000	1.000	1.000	1.000	1.000	1.000
Numérica07**	300	0.823	0.412*+	0.838	0.810	0.860	0.782	0.815
Numérica08**	300	0.709	0.094*+	0.150*+	0.408*+	0.381*+	0.571	0.580
Numérica09**	300	0.485	0.067*+	0.192*+	0.274*+	0.248*+	0.598	0.615
Numérica10**	300	0.552	0.002*+	0.182*+	0.257*+	0.267*+	0.614	0.634
Combinada01**	900	0.510	0.900+	0.900+	0.900+	0.910+	0.900+	1.000*
Combinada02**	900	0.000	0.640*+	0.680*+	0.250*+	0.270*+	0.730	0.740
Combinada03**	900	0.200	0.610+	0.670+	0.430*+	0.530+	0.660+	0.810*
Combinada04	900	0.210	0.720	0.730	0.730	0.750	0.710	0.750
Combinada05**	900	0.260	0.645	0.570	0.505*+	0.560	0.615	0.625

Tabla 3: Resultados del recuerdo con BD artificiales y clasificador K-NN.

Nombre	**%sob**	**Original**	**ROS**	**SMOTE**	**BSM1**	**BSM2**	**GIS-C**	**GIS-CJ**
Nominal01**	900	0.000	0.490	0.440	0.030*+	0.030*+	0.500	0.500
Nominal02**	900	0.000	0.320	0.200*+	0.000*+	0.000*+	0.330	0.330
Nominal03**	400	0.000	0.530	0.430*+	0.035*+	0.035*+	0.515	0.515
Nominal04**	400	0.000	0.520	0.440*+	0.015*+	0.040*+	0.520	0.520
Nominal05**	400	0.000	0.622*+	0.482*+	0.010*+	0.010*+	0.624	0.624
Numérica01	800	0.900	1.000	1.000	1.000	1.000	1.000	1.000
Numérica02**	800	0.000	0.185*+	0.695+	0.605*+	0.675+	0.690+	0.750*
Numérica03**	800	0.008	0.015*+	0.220*+	0.322*+	0.348*+	0.575	0.588
Numérica04**	800	0.000	0.038*+	0.057*+	0.180*+	0.173*+	0.248	0.235
Numérica05**	800	0.000	0.000*+	0.732*+	0.573	0.517	0.485	0.525
Numérica06	300	0.950	1.000	1.000	1.000	1.000	1.000	1.000
Numérica07**	300	0.000	0.412*+	0.838	0.810	0.860	0.782	0.815
Numérica08**	300	0.261	0.094*+	0.150*+	0.408*+	0.381*+	0.571	0.580
Numérica09**	300	0.000	0.067*+	0.192*+	0.274*+	0.248*+	0.598	0.615
Numérica10**	300	0.000	0.002*+	0.182*+	0.257*+	0.267*+	0.614	0.634
Combinada01**	900	0.900	0.900+	0.900+	0.900+	0.900+	0.900+	1.000*
Combinada02**	900	0.000	0.640*+	0.680*+	0.250*+	0.270*+	0.730	0.740
Combinada03**	900	0.210	0.610+	0.670+	0.430*+	0.530+	0.660+	0.810*
Combinada04	900	0.470	0.720	0.730	0.730	0.750	0.710	0.750
Combinada05**	900	0.290	0.645	0.570	0.505	0.560	0.615	0.625

Tabla 4: Resultados del recuerdo con BD artificiales y clasificador C4.5.

Nombre	**%sob**	**Original**	**ROS**	**SMOTE**	**BSM1**	**BSM2**	**GIS-C**	**GIS-CJ**
Nominal01**	900	0.000	0.490	0.440*+	0.030*+	0.030*+	0.500	0.500
Nominal02**	900	0.000	0.320	0.200*+	0.000*+	0.000*+	0.330	0.330
Nominal03**	400	0.000	0.530	0.430*+	0.035*+	0.035*+	0.515	0.515
Nominal04**	400	0.000	0.520	0.440*+	0.015*+	0.040*+	0.520	0.520
Nominal05**	400	0.000	0.622	0.482*+	0.010*+	0.010*+	0.624	0.624
Numérica01	800	0.900	1.000	1.000	1.000	1.000	1.000	1.000
Numérica02**	800	0.000	0.185*+	0.695+	0.605+	0.675+	0.690+	0.750*
Numérica03**	800	0.000	0.015*+	0.220*+	0.322*+	0.348*+	0.575	0.588
Numérica04**	800	0.000	0.038*+	0.057*+	0.180*+	0.173*+	0.248	0.235
Numérica05**	800	0.000	0.000*+	0.732*+	0.573	0.517	0.485	0.525
Numérica06	300	0.950	1.000	1.000	1.000	1.000	1.000	1.000
Numérica07**	300	0.000	0.412*+	0.838	0.810	0.860	0.782	0.815
Numérica08**	300	0.138	0.094*+	0.150*+	0.408*+	0.381*+	0.571	0.580
Numérica09**	300	0.000	0.067*+	0.192*+	0.274*+	0.248*+	0.598	0.615
Numérica10**	300	0.000	0.002*+	0.182*+	0.257*+	0.267*+	0.614	0.634
Combinada01**	900	0.900	0.900+	0.900+	0.900+	0.900+	0.900+	1.000*
Combinada02**	900	0.000	0.640*+	0.680*+	0.250*+	0.270*+	0.730	0.740
Combinada03**	900	0.210	0.610+	0.670+	0.430*+	0.530*+	0.660+	0.810*
Combinada04	900	0.470	0.720	0.730	0.730	0.750	0.710	0.750
Combinada05*	900	0.305	0.645	0.570	0.505	0.560	0.615	0.625

Tabla 5: Resultados del recuerdo con BD artificiales y clasificador PART.

Nombre	**%sob**	**Original**	**ROS**	**SMOTE**	**BSM1**	**BSM2**	**GIS-C**	**GIS-CJ**
Nominal01**	900	0.000	0.490	0.440	0.030*+	0.030*+	0.500	0.500
Nominal02**	900	0.000	0.320	0.200*+	0.000*+	0.000*+	0.330	0.330
Nominal03**	400	0.000	0.530	0.430*+	0.035*+	0.035*+	0.515	0.515
Nominal04**	400	0.000	0.520	0.440*+	0.015*+	0.040*+	0.520	0.520
Nominal05**	400	0.010	0.622	0.482*+	0.010*+	0.010*+	0.624	0.624
Numérica01	800	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Numérica02**	800	0.115	0.185*+	0.695	0.605+	0.675	0.690	0.750
Numérica03**	800	0.000	0.015*+	0.220*+	0.322*+	0.348*+	0.575	0.588
Numérica04**	800	0.000	0.038*+	0.057*+	0.180*+	0.173*+	0.248	0.235
Numérica05**	800	0.000	0.000*+	0.732*+	0.573*+	0.517	0.485	0.525
Numérica06	300	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Numérica07**	300	0.400	0.412*+	0.838	0.810	0.860	0.782	0.815
Numérica08**	300	0.074	0.094*+	0.150*+	0.408*+	0.381*+	0.571	0.580
Numérica09**	300	0.038	0.067*+	0.192*+	0.274*+	0.248*+	0.598	0.615
Numérica10**	300	0.000	0.002*+	0.182*+	0.257*+	0.267*+	0.614	0.634
Combinada01**	900	0.900	0.900+	0.900+	0.900+	0.910+	0.900+	1.000*
Combinada02**	900	0.000	0.640*+	0.680*+	0.250*+	0.270*+	0.730	0.740
Combinada03**	900	0.620	0.610+	0.670+	0.430*+	0.530*+	0.660+	0.810*
Combinada04	900	0.250	0.720	0.730	0.730	0.750	0.710	0.750
Combinada05**	900	0.030	0.645	0.570	0.505	0.560	0.615	0.625

Tabla 6: Resultados del recuerdo con BD artificiales y clasificador *Backpropagation*.

APÉNDICE B

Resultados de la precisión sobre bases de datos artificiales

Este apéndice contiene las tablas de resultados de la precisión de los seis métodos de sobre-muestreo (ROS, SMOTE, BSM1, BSM2, GIS-G y GIS-GF) con las veinte bases de datos generadas artificialmente. Son seis tablas, cada una de ellas corresponde a cada clasificador utilizado: *AdaBoost M1*, *Naive Bayes*, K-NN, C4.5, PART y *Backpropagation*.

Una marca ** delante del nombre de la base de datos indica que ANOVA mostró que los resultados en general son significativos, no se deben a la aleatoriedad introducida por los métodos de sobre-muestreo. Un signo * delante de cada valor del recuerdo indica que la diferencia de ese valor con el obtenido por GIS-G es significativa y un signo + que lo es con GIS-GF.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Nominal01**	900	0.000	0.098*+	0.126*+	0.009*+	0.009*+	0.089	0.089
Nominal02**	900	0.000	0.053	0.070	0.000	0.000	0.052	0.052
Nominal03**	400	0.000	0.217	0.209	0.021*+	0.017*+	0.205	0.205
Nominal04**	400	0.000	0.201	0.208	0.009*+	0.022*+	0.190	0.190
Nominal05**	400	0.000	0.166	0.146	0.017*+	0.017*+	0.161	0.161
Numérica01	800	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Numérica02**	800	0.000	0.338*+	0.303*+	0.398*+	0.379*+	0.492+	0.415*
Numérica03**	800	0.000	0.043*+	0.059*+	0.131	0.137	0.117	0.105
Numérica04**	800	0.000	0.132	0.080*+	0.097*+	0.114	0.106	0.128
Numérica05**	800	0.000	0.000*+	0.088	0.100*+	0.092	0.093	0.089
Numérica06**	300	1.000	0.995	0.995	0.995	0.920	0.995	0.995
Numérica07**	300	0.000	0.623*+	0.596*+	0.539*+	0.530*+	0.633+	0.619*
Numérica08**	300	0.000	0.285	0.155*+	0.330*+	0.317*+	0.247	0.244
Numérica09	300	0.000	0.314	0.285	0.330*+	0.343*+	0.280	0.280
Numérica10**	300	0.000	0.020*+	0.176	0.117	0.144	0.157	0.158
Combinada01**	900	1.000	0.860+	0.855	0.860+	0.813*+	0.873+	0.932
Combinada02**	900	0.000	0.133*+	0.153*+	0.145*+	0.122*+	0.159+	0.173*
Combinada03**	900	0.900	0.560	0.595	0.388	0.428	0.499	0.471
Combinada04	900	1.000	0.303	0.302	0.361	0.321	0.348	0.297
Combinada05**	900	0.990	0.257	0.364	0.309	0.262	0.306	0.224

Tabla 7: Resultados Precisión con BD artificiales y clasificador AdaBoostM1.

Nombre	**%sob**	**Original**	**ROS**	**SMOTE**	**BSM1**	**BSM2**	**GIS-C**	**GIS-CJ**
Nominal01**	900	0.000	0.098	0.126*+	0.009	0.009	0.089	0.089
Nominal02**	900	0.000	0.053	0.070	0.000	0.000	0.052	0.052
Nominal03**	400	0.000	0.217	0.209	0.021	0.017	0.205	0.205
Nominal04**	400	0.000	0.201	0.208	0.009*+	0.022*+	0.190	0.190
Nominal05**	400	0.000	0.166	0.146	0.017*+	0.017*+	0.161	0.161
Numérica01**	800	0.800	1.000	1.000	1.000	0.893*+	1.000	1.000
Numérica02**	800	0.000	0.338*+	0.303*+	0.398*+	0.379*+	0.492+	0.415*
Numérica03**	800	0.000	0.043*+	0.059*+	0.131*+	0.137*+	0.117	0.105
Numérica04**	800	0.000	0.132	0.080	0.097	0.114	0.106	0.128
Numérica05**	800	0.000	0.000*+	0.088	0.100	0.092	0.093	0.089
Numérica06**	300	1.000	0.995	0.995	0.995	0.920	0.995	0.995
Numérica07	300	0.000	0.623	0.596	0.539	0.530	0.633	0.619
Numérica08**	300	0.000	0.285	0.155*+	0.330*+	0.317*+	0.247	0.244
Numérica09	300	0.000	0.314	0.285	0.330	0.343	0.280	0.280
Numérica10**	300	0.000	0.020	0.176	0.117	0.144	0.157	0.158
Combinada01**	900	0.700	0.860+	0.855	0.860+	0.813*+	0.873+	0.932*
Combinada02**	900	0.000	0.133*+	0.153*+	0.145*+	0.122*+	0.159+	0.173*
Combinada03**	900	0.000	0.560*+	0.595*+	0.388*+	0.428	0.499+	0.471*
Combinada04	900	0.000	0.303	0.302	0.361	0.321	0.348	0.297
Combinada05**	900	0.000	0.257	0.364	0.309	0.262	0.306	0.224

Tabla 8: Resultados de la precisión con BD artificiales y clasificador *Naive Bayes*.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Nominal01**	900	0.000	0.098	0.126	0.009	0.009	0.089	0.089
Nominal02**	900	0.000	0.053	0.070	0.000	0.000	0.052	0.052
Nominal03**	400	0.000	0.217	0.209	0.021*+	0.017*+	0.205	0.205
Nominal04**	400	0.000	0.201	0.208	0.009*+	0.022*+	0.190	0.190
Nominal05**	400	0.002	0.166	0.146	0.017*+	0.017*+	0.161	0.161
Numérica01**	800	0.547	1.000	1.000	1.000	0.893*+	1.000	1.000
Numérica02**	800	0.930	0.338	0.303	0.398	0.379	0.492	0.415
Numérica03**	800	0.920	0.043*+	0.059*+	0.131	0.137	0.117	0.105
Numérica04	800	0.870	0.132	0.080+	0.097+	0.114	0.106	0.128
Numérica05**	800	0.730	0.000*+	0.088	0.100	0.092	0.093	0.089
Numérica06**	300	0.980	0.995	0.995	0.995	0.920	0.995	0.995
Numérica07	300	0.993	0.623	0.596	0.539	0.530	0.633	0.619
Numérica08**	300	0.903	0.285	0.155*+	0.330*+	0.317*+	0.247	0.244
Numérica09	300	0.895	0.314	0.285	0.330	0.343	0.280	0.280
Numérica10**	300	0.784	0.020*+	0.176	0.117	0.144	0.157	0.158
Combinada01**	900	0.488	0.860+	0.855+	0.860+	0.813+	0.873+	0.932*
Combinada02**	900	0.000	0.133*+	0.153	0.145	0.122*+	0.159	0.173
Combinada03**	900	0.190	0.560	0.595	0.388*+	0.428	0.499	0.471
Combinada04	900	0.210	0.303	0.302	0.361	0.321	0.348	0.297
Combinada05**	900	0.359	0.257	0.364	0.309	0.262	0.306	0.224

Tabla 9: Resultados de la precisión con BD artificiales y clasificador K-NN.

Nombre	**%sob**	**Original**	**ROS**	**SMOTE**	**BSM1**	**BSM2**	**GIS-C**	**GIS-CJ**
Nominal01**	900	0.000	0.098	0.126	0.009	0.009	0.089	0.089
Nominal02**	900	0.000	0.053	0.070	0.000	0.000	0.052	0.052
Nominal03**	400	0.000	0.217	0.209	0.021*+	0.017*+	0.205	0.205
Nominal04**	400	0.000	0.201	0.208	0.009*+	0.022*+	0.190	0.190
Nominal05**	400	0.000	0.166	0.146	0.017*+	0.017*+	0.161	0.161
Numérica01**	800	0.900	1.000	1.000	1.000	0.893*+	1.000	0.995
Numérica02**	800	0.000	0.338*+	0.303*+	0.398*+	0.379*+	0.492+	0.415*
Numérica03**	800	0.020	0.043*+	0.059*+	0.131*+	0.137*+	0.117	0.105
Numérica04**	800	0.000	0.132	0.080*+	0.097*+	0.114	0.106	0.128
Numérica05**	800	0.000	0.000*+	0.088	0.100	0.092	0.093	0.089
Numérica06**	300	1.000	0.995	0.995	0.995	0.920	0.995	0.995
Numérica07	300	0.000	0.623	0.596	0.539	0.530	0.633	0.619
Numérica08**	300	0.298	0.285	0.155*+	0.330*+	0.317*+	0.247	0.244
Numérica09	300	0.000	0.314	0.285	0.330	0.343	0.280	0.280
Numérica10**	300	0.000	0.020*+	0.176	0.117	0.144	0.157	0.158
Combinada01**	900	0.900	0.860	0.855	0.860	0.813	0.873	0.932
Combinada02**	900	0.000	0.133	0.153	0.145	0.122	0.159	0.173
Combinada03**	900	0.210	0.560	0.595	0.388	0.428	0.499	0.471
Combinada04	900	0.410	0.303	0.302	0.361	0.321	0.348	0.297
Combinada05**	900	0.371	0.257	0.364	0.309	0.262	0.306	0.224

Tabla 10: Resultados de la precisión con BD artificiales y clasificador C4.5.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Nominal01**	900	0.000	0.098	0.126	0.009	0.009	0.089	0.089
Nominal02**	900	0.000	0.053	0.070	0.000	0.000	0.052	0.052
Nominal03**	400	0.000	0.217	0.209	0.021*+	0.017*+	0.205	0.205
Nominal04**	400	0.000	0.201	0.208	0.009*+	0.022*+	0.190	0.190
Nominal05**	400	0.000	0.166	0.146	0.017*+	0.017*+	0.161	0.161
Numérica01**	800	0.900	1.000	1.000	1.000	0.893*+	1.000	1.000
Numérica02**	800	0.000	0.338*+	0.303*+	0.398*+	0.379*+	0.492*	0.415+
Numérica03**	800	0.000	0.043*+	0.059*+	0.131	0.137	0.117	0.105
Numérica04**	800	0.000	0.132	0.080+	0.097+	0.114	0.106	0.128
Numérica05**	800	0.000	0.000*+	0.088	0.100	0.092	0.093	0.089
Numérica06**	300	1.000	0.995	0.995	0.995	0.920	0.995	0.995
Numérica07	300	0.000	0.623	0.596	0.539	0.530	0.633	0.619
Numérica08**	300	0.215	0.285	0.155*+	0.330*+	0.317*+	0.247	0.244
Numérica09	300	0.000	0.314	0.285	0.330	0.343	0.280	0.280
Numérica10**	300	0.000	0.020*+	0.176	0.117	0.144	0.157	0.158
Combinada01**	900	0.860	0.860	0.855	0.860	0.813+	0.873	0.932
Combinada02**	900	0.000	0.133	0.153	0.145	0.122	0.159	0.173
Combinada03**	900	0.210	0.560	0.595	0.388*+	0.428	0.499	0.471
Combinada04	900	0.410	0.303	0.302	0.361	0.321	0.348	0.297
Combinada05**	900	0.401	0.257	0.364	0.309	0.262	0.306	0.224

Tabla 11: Resultados de la precisión con BD artificiales y clasificador PART.

Nombre	**%sob**	**Original**	**ROS**	**SMOTE**	**BSM1**	**BSM2**	**GIS-C**	**GIS-CJ**
Nominal01**	900	0.000	0.098	0.126	0.009	0.009	0.089	0.089
Nominal02**	900	0.000	0.053	0.070	0.000	0.000	0.052	0.052
Nominal03**	400	0.000	0.217	0.209	0.021*+	0.017*+	0.205	0.205
Nominal04**	400	0.000	0.201	0.208	0.009*+	0.022*+	0.190	0.190
Nominal05**	400	0.017	0.166	0.146	0.017*+	0.017*+	0.161	0.161
Numérica01**	800	1.000	1.000	1.000	1.000	0.893*+	1.000	1.000
Numérica02**	800	0.220	0.338	0.303*+	0.398	0.379	0.492	0.415
Numérica03**	800	0.000	0.043*+	0.059*+	0.131	0.137	0.117	0.105
Numérica04**	800	0.000	0.132	0.080	0.097	0.114	0.106	0.128
Numérica05**	800	0.000	0.000*+	0.088	0.100	0.092	0.093	0.089
Numérica06**	300	0.995	0.995	0.995	0.995	0.920	0.995	0.995
Numérica07	300	0.561	0.623	0.596	0.539	0.530	0.633	0.619
Numérica08**	300	0.232	0.285	0.155*+	0.330*+	0.317*+	0.247	0.244
Numérica09	300	0.175	0.314	0.285	0.330	0.343	0.280	0.280
Numérica10**	300	0.000	0.020	0.176	0.117	0.144	0.157	0.158
Combinada01**	900	0.895	0.860	0.855	0.860	0.813	0.873	0.932
Combinada02**	900	0.000	0.133	0.153	0.145	0.122	0.159	0.173
Combinada03**	900	0.573	0.560	0.595	0.388	0.428	0.499	0.471
Combinada04	900	0.250	0.303	0.302	0.361	0.321	0.348	0.297
Combinada05**	900	0.052	0.257	0.364	0.309	0.262	0.306	0.224

Tabla 12: Resultados de la precisión con BD artificiales y clasificador *Backpropagation*.

APÉNDICE C

Resultados de la Medida-F sobre bases de datos artificiales

Este apéndice contiene las tablas de resultados de la Medida-F de los seis métodos de sobre-muestreo (ROS, SMOTE, BSM1, BSM2, GIS-G y GIS-GF) con las veinte bases de datos generadas artificialmente. Son seis tablas, cada una de ellas corresponde a cada clasificador utilizado: *AdaBoost M1*, *Naive Bayes*, K-NN, C4.5, PART y *Backpropagation*.

Una marca ** delante del nombre de la base de datos indica que ANOVA mostró que los resultados en general son significativos, no se deben a la aleatoriedad introducida por los métodos de sobre-muestreo. Un signo * delante de cada valor del recuerdo indica que la diferencia de ese valor con el obtenido por GIS-G es significativa y un signo + que lo es con GIS-GF.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Nominal01**	900	0.000	0.162	0.190	0.014*+	0.014*+	0.150	0.150
Nominal02**	900	0.000	0.090	0.094	0.000*+	0.000*+	0.089	0.089
Nominal03**	400	0.000	0.300*+	0.271*+	0.026*+	0.023*+	0.282	0.282
Nominal04**	400	0.000	0.279	0.268	0.011*+	0.028*+	0.273	0.273
Nominal05**	400	0.000	0.260	0.222	0.012*+	0.012*+	0.254	0.254
Numérica01	800	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Numérica02**	800	0.000	0.236*+	0.399*+	0.442*	0.456*	0.540+	0.487*
Numérica03**	800	0.000	0.022*+	0.088*+	0.175	0.186	0.171	0.166
Numérica04**	800	0.000	0.057	0.050	0.101	0.109	0.093	0.093
Numérica05**	800	0.000	0.000*+	0.156*+	0.162*+	0.148*+	0.120	0.127
Numérica06**	300	1.000	0.997	0.997	0.997	0.950	0.997	0.997
Numérica07**	300	0.000	0.459*+	0.674	0.618	0.628	0.673	0.674
Numérica08**	300	0.000	0.131*+	0.141*+	0.345	0.314	0.304	0.313
Numérica09**	300	0.000	0.108*+	0.159*+	0.203*+	0.210*+	0.283	0.287
Numérica10**	300	0.000	0.003*+	0.137*+	0.145*+	0.154*+	0.228	0.236
Combinada01**	900	0.900	0.873+	0.870+	0.873+	0.845+	0.882+	0.954*
Combinada02**	900	0.000	0.216*+	0.240+	0.172*+	0.157*+	0.255+	0.271*
Combinada03**	900	0.900	0.577	0.618*+	0.402*+	0.460*+	0.547	0.568
Combinada04	900	1.000	0.396*	0.403	0.455+	0.420	0.441+	0.391*
Combinada05**	900	0.837	0.340	0.426*+	0.357	0.336	0.385	0.308

Tabla 13: Resultados de la Medida-F con BD artificiales y clasificador AdaBoostM1.

Nombre	**%sob**	**Original**	**ROS**	**SMOTE**	**BSM1**	**BSM2**	**GIS-C**	**GIS-CJ**
Nominal01**	900	0.000	0.162	0.190*+	0.014*+	0.014*+	0.150	0.150
Nominal02**	900	0.000	0.090	0.094	0.000*+	0.000*+	0.089	0.089
Nominal03**	400	0.000	0.300	0.271	0.026*+	0.023*+	0.282	0.282
Nominal04**	400	0.000	0.279	0.268	0.011*+	0.028*+	0.273	0.273
Nominal05**	400	0.000	0.260	0.222*+	0.012*+	0.012*+	0.254	0.254
Numérica01**	800	0.800	1.000	1.000	1.000	0.927	1.000	1.000
Numérica02**	800	0.000	0.236*+	0.399*+	0.442*	0.456*	0.540+	0.487*
Numérica03**	800	0.000	0.022*+	0.088*+	0.175	0.186	0.171+	0.166*
Numérica04**	800	0.000	0.057*+	0.050*+	0.101	0.109	0.093	0.093
Numérica05**	800	0.000	0.000*+	0.156	0.162	0.148	0.120	0.127
Numérica06**	300	0.933	0.997	0.997	0.997	0.950	0.997	0.997
Numérica07**	300	0.000	0.459*+	0.674	0.618	0.628	0.673	0.674
Numérica08**	300	0.000	0.131*+	0.141*+	0.345	0.314	0.304	0.313
Numérica09**	300	0.000	0.108*+	0.159*+	0.203*+	0.210*+	0.283	0.287
Numérica10**	300	0.000	0.003*+	0.137*+	0.145*+	0.154*+	0.228	0.236
Combinada01**	900	0.700	0.873+	0.870+	0.873+	0.845+	0.882+	0.954*
Combinada02**	900	0.000	0.216	0.240	0.172*+	0.157*+	0.255+	0.271*
Combinada03**	900	0.000	0.577	0.618*+	0.402*+	0.460*+	0.547+	0.568*
Combinada04**	900	0.000	0.396*	0.403	0.455*	0.420*	0.441+	0.391*
Combinada05**	900	0.000	0.340*+	0.426*+	0.357*+	0.336*+	0.385+	0.308*

Tabla 14: Resultados de la Medida-F con BD artificiales y clasificador *Naive Bayes*.

Nombre	**%sob**	**Original**	**ROS**	**SMOTE**	**BSM1**	**BSM2**	**GIS-C**	**GIS-CJ**
Nominal01**	900	0.000	0.162	0.190	0.014*+	0.014*+	0.150	0.150
Nominal02**	900	0.000	0.090	0.094	0.000*+	0.000*+	0.089	0.089
Nominal03**	400	0.000	0.300	0.271	0.026*+	0.023*+	0.282	0.282
Nominal04**	400	0.000	0.279	0.268	0.011*+	0.028*+	0.273	0.273
Nominal05**	400	0.002	0.260	0.222*+	0.012*+	0.012*+	0.254	0.254
Numérica01**	800	0.563	1.000	1.000	1.000	0.927	1.000	1.000
Numérica02**	800	0.787	0.236*+	0.399*+	0.442	0.456	0.540	0.487
Numérica03**	800	0.588	0.022*+	0.088*+	0.175	0.186	0.171	0.166
Numérica04**	800	0.483	0.057	0.050	0.101	0.109	0.093	0.093
Numérica05**	800	0.335	0.000*+	0.156	0.162	0.148	0.120	0.127
Numérica06**	300	0.925	0.997	0.997	0.997	0.950	0.997	0.997
Numérica07**	300	0.890	0.459*+	0.674	0.618	0.628	0.673	0.674
Numérica08**	300	0.779	0.131*+	0.141*+	0.345	0.314	0.304	0.313
Numérica09**	300	0.612	0.108*+	0.159*+	0.203*+	0.210*+	0.283	0.287
Numérica10**	300	0.635	0.003*+	0.137*+	0.145*+	0.154*+	0.228	0.236
Combinada01**	900	0.492	0.873+	0.870+	0.873+	0.845+	0.882+	0.954*
Combinada02**	900	0.000	0.216	0.240	0.172*+	0.157*+	0.255	0.271
Combinada03**	900	0.193	0.577	0.618	0.402*+	0.460	0.547	0.568
Combinada04	900	0.210	0.396	0.403	0.455	0.420	0.441	0.391
Combinada05**	900	0.289	0.340	0.426	0.357	0.336	0.385	0.308

Tabla 15: Resultados de la Medida-F con BD artificiales y clasificador K-NN.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Nominal01**	900	0.000	0.162*+	0.190*+	0.014*+	0.014*+	0.150	0.150
Nominal02**	900	0.000	0.090	0.094	0.000*+	0.000*+	0.089	0.089
Nominal03**	400	0.000	0.300	0.271	0.026*+	0.023*+	0.282	0.282
Nominal04**	400	0.000	0.279	0.268	0.011*+	0.028*+	0.273	0.273
Nominal05**	400	0.000	0.260	0.222*+	0.012*+	0.012*+	0.254	0.254
Numérica01**	800	0.900	1.000	1.000	1.000	0.927	1.000	1.000
Numérica02**	800	0.000	0.236*+	0.399*+	0.442*	0.456*	0.540+	0.487*
Numérica03**	800	0.011	0.022*+	0.088*+	0.175	0.186	0.171	0.166
Numérica04**	800	0.000	0.057	0.050	0.101	0.109	0.093	0.093
Numérica05**	800	0.000	0.000*+	0.156	0.162	0.148	0.120	0.127
Numérica06**	300	0.967	0.997	0.997	0.997	0.950	0.997	0.997
Numérica07**	300	0.000	0.459*+	0.674	0.618	0.628	0.673	0.674
Numérica08**	300	0.271	0.131*+	0.141*+	0.345	0.314	0.304	0.313
Numérica09**	300	0.000	0.108*+	0.159*+	0.203*+	0.210*+	0.283	0.287
Numérica10**	300	0.000	0.003*+	0.137*+	0.145*+	0.154*+	0.228	0.236
Combinada01**	900	0.900	0.873+	0.870+	0.873+	0.845+	0.882+	0.954*
Combinada02**	900	0.000	0.216	0.240	0.172*+	0.157*+	0.255	0.271
Combinada03**	900	0.210	0.577	0.618*+	0.402*+	0.460*+	0.547	0.568
Combinada04	900	0.428	0.396	0.403	0.455	0.420	0.441	0.391
Combinada05**	900	0.312	0.340	0.426	0.357	0.336	0.385	0.308

Tabla 16: Resultados de la Medida-F con BD artificiales y clasificador C4.5.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Nominal01**	900	0.000	0.162	0.190	0.014*+	0.014*+	0.150	0.150
Nominal02**	900	0.000	0.090	0.094	0.000*+	0.000*+	0.089	0.089
Nominal03**	400	0.000	0.300	0.271	0.026*+	0.023*+	0.282	0.282
Nominal04**	400	0.000	0.279	0.268	0.011*+	0.028*+	0.273	0.273
Nominal05**	400	0.000	0.260	0.222	0.012*+	0.012*+	0.254	0.254
Numérica01**	800	0.900	1.000	1.000	1.000	0.927	1.000	1.000
Numérica02**	800	0.000	0.236+*	0.399+*	0.442*	0.456*	0.540+	0.487*
Numérica03**	800	0.000	0.022+*	0.088+*	0.175	0.186	0.171	0.166
Numérica04**	800	0.000	0.057	0.050	0.101	0.109	0.093	0.093
Numérica05**	800	0.000	0.000+*	0.156	0.162	0.148	0.120	0.127
Numérica06**	300	0.967	0.997	0.997	0.997	0.950	0.997	0.997
Numérica07**	300	0.000	0.459+*	0.674	0.618	0.628	0.673	0.674
Numérica08**	300	0.158	0.131+*	0.141+*	0.345	0.314	0.304	0.313
Numérica09**	300	0.000	0.108+*	0.159+*	0.203+*	0.210+*	0.283	0.287
Numérica10**	300	0.000	0.003+*	0.137+*	0.145+*	0.154+*	0.228	0.236
Combinada01**	900	0.900	0.873	0.870	0.873	0.845	0.882	0.954
Combinada02**	900	0.000	0.216	0.240	0.172*+	0.157*+	0.255	0.271
Combinada03**	900	0.210	0.577	0.618	0.402+	0.460*+	0.547	0.568
Combinada04	900	0.428	0.396	0.403	0.455	0.420	0.441	0.391
Combinada05**	900	0.332	0.340	0.426	0.357	0.336	0.385	0.308

Tabla 17: Resultados de la Medida-F con BD artificiales y clasificador PART.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Nominal01**	900	0.000	0.162	0.190*+	0.014*+	0.014*+	0.150	0.150
Nominal02**	900	0.000	0.090	0.094	0.000*+	0.000*+	0.089	0.089
Nominal03**	400	0.000	0.300	0.271	0.026*+	0.023*+	0.282	0.282
Nominal04**	400	0.000	0.279	0.268	0.011*+	0.028*+	0.273	0.273
Nominal05**	400	0.012	0.260	0.222	0.012*+	0.012*+	0.254	0.254
Numérica01**	800	1.000	1.000	1.000	1.000	0.927	1.000	1.000
Numérica02**	800	0.150	0.236*+	0.399*+	0.442	0.456	0.540	0.487
Numérica03**	800	0.000	0.022*+	0.088*+	0.175	0.186	0.171	0.166
Numérica04**	800	0.000	0.057	0.050	0.101	0.109	0.093	0.093
Numérica05**	800	0.000	0.000*+	0.156	0.162	0.148	0.120	0.127
Numérica06**	300	0.997	0.997	0.997	0.997	0.950	0.997	0.997
Numérica07**	300	0.433	0.459*+	0.674	0.618	0.628	0.673	0.674
Numérica08**	300	0.106	0.131*+	0.141*+	0.345	0.314	0.304	0.313
Numérica09**	300	0.060	0.108*+	0.159*+	0.203*+	0.210*+	0.283	0.287
Numérica10**	300	0.000	0.003*+	0.137*+	0.145*+	0.154*+	0.228	0.236
Combinada01**	900	0.897	0.873	0.870	0.873	0.845	0.882	0.954
Combinada02**	900	0.000	0.216	0.240	0.172*+	0.157*+	0.255	0.271
Combinada03**	900	0.588	0.577	0.618	0.402	0.460	0.547	0.568
Combinada04	900	0.250	0.396	0.403	0.455	0.420	0.441	0.391
Combinada05**	900	0.036	0.340*+	0.426*+	0.357*+	0.336*+	0.385+	0.308*

Tabla 18: Resultados de la Medida-F con BD artificiales y clasificador *Backpropagation*.

APÉNDICE D

Resultados del AUC sobre bases de datos artificiales

Este apéndice contiene las tablas de resultados del AUC de los seis métodos de sobre-muestreo (ROS, SMOTE, BSM1, BSM2, GIS-G y GIS-GF) con las veinte bases de datos generadas artificialmente. Son seis tablas, cada una de ellas corresponde a cada clasificador utilizado: *AdaBoost M1*, *Naive Bayes*, K-NN, C4.5, PART y *Backpropagation*.

Una marca ** delante del nombre de la base de datos indica que ANOVA mostró que los resultados en general son significativos, no se deben a la aleatoriedad introducida por los métodos de sobre-muestreo. Un signo * delante de cada valor del recuerdo indica que la diferencia de ese valor con el obtenido por GIS-G es significativa y un signo + que lo es con GIS-GF.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Nominal01**	900	0.547	0.541	0.602	0.636	0.634	0.532	0.532
Nominal02**	900	0.468	0.405	0.420	0.471	0.471	0.404	0.404
Nominal03**	400	0.580	0.577	0.565	0.563	0.556	0.590	0.590
Nominal04	400	0.477	0.499	0.522	0.518	0.535	0.497	0.497
Nominal05	400	0.543	0.526	0.518	0.523	0.523	0.537	0.537
Numérica01	800	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Numérica02**	800	0.444	0.361*+	0.798	0.708	0.725	0.845	0.874
Numérica03	800	0.591	0.555	0.494	0.569	0.579	0.508	0.520
Numérica04	800	0.427	0.496	0.456	0.440	0.459	0.467	0.452
Numérica05	800	0.462	0.417	0.414	0.486	0.460	0.427	0.443
Numérica06	300	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Numérica07	300	0.445	0.569*+	0.893	0.832	0.859	0.878	0.870
Numérica08	300	0.708	0.567	0.482	0.608	0.587	0.483	0.498
Numérica09	300	0.407	0.517	0.475	0.529	0.518	0.505	0.540
Numérica10	300	0.453	0.474	0.459	0.458	0.470	0.445	0.462
Combinada01**	900	0.900	0.998	0.997	0.999	0.998	0.997	0.999
Combinada02**	900	0.741	0.702	0.672	0.754	0.736	0.699	0.724
Combinada03**	900	0.956	0.970	0.897	0.692*+	0.772*+	0.953	0.930
Combinada04	900	1.000	0.741	0.744	0.757	0.753	0.764	0.759
Combinada05	900	0.965	0.748	0.684	0.686	0.700	0.716	0.722

Tabla 19: Resultados del AUC con BD artificiales y clasificador AdaBoostM1.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Nominal01**	900	0.579	0.541	0.602	0.636	0.634	0.532	0.532
Nominal02**	900	0.431	0.405	0.420	0.471	0.471	0.404	0.404
Nominal03	400	0.565	0.577	0.565	0.563	0.556	0.590	0.590
Nominal04	400	0.508	0.499	0.522	0.518	0.535	0.497	0.497
Nominal05	400	0.541	0.526	0.518	0.523	0.523	0.537	0.537
Numérica01	800	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Numérica02**	800	0.229	0.361*+	0.798	0.708*+	0.725*+	0.845	0.874
Numérica03**	800	0.645	0.555	0.494	0.569	0.579	0.508	0.520
Numérica04**	800	0.295	0.496	0.456	0.440	0.459	0.467	0.452
Numérica05**	800	0.491	0.417	0.414	0.486	0.460	0.427	0.443
Numérica06	300	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Numérica07**	300	0.300	0.569*+	0.893	0.832	0.859	0.878	0.870
Numérica08**	300	0.562	0.567	0.482	0.608	0.587	0.483	0.498
Numérica09	300	0.346	0.517	0.475	0.529	0.518	0.505	0.540
Numérica10	300	0.490	0.474	0.459	0.458	0.470	0.445	0.462
Combinada01	900	0.900	0.998	0.997	0.999	0.998	0.997	0.999
Combinada02	900	0.730	0.702	0.672	0.754	0.736	0.699	0.724
Combinada03**	900	0.971	0.970	0.897	0.692*+	0.772*+	0.953	0.930
Combinada04**	900	0.715	0.741	0.744	0.757	0.753	0.764	0.759
Combinada05	900	0.690	0.748	0.684	0.686	0.700	0.716	0.722

Tabla 20: Resultados del AUC con BD artificiales y clasificador *Naive Bayes*.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Nominal01**	900	0.522	0.541	0.602	0.636	0.634	0.532	0.532
Nominal02**	900	0.394	0.405	0.420	0.471	0.471	0.404	0.404
Nominal03	400	0.580	0.577	0.565	0.563	0.556	0.590	0.590
Nominal04	400	0.498	0.499	0.522	0.518	0.535	0.497	0.497
Nominal05	400	0.536	0.526	0.518	0.523	0.523	0.537	0.537
Numérica01	800	0.990	1.000	1.000	1.000	1.000	1.000	1.000
Numérica02**	800	0.871	0.361*+	0.798	0.708*+	0.725*+	0.845	0.874
Numérica03**	800	0.909	0.555	0.494	0.569	0.579	0.508	0.520
Numérica04	800	0.823	0.496	0.456	0.440	0.459	0.467	0.452
Numérica05**	800	0.759	0.417	0.414	0.486	0.460	0.427	0.443
Numérica06	300	0.996	1.000	1.000	1.000	1.000	1.000	1.000
Numérica07**	300	0.969	0.569*+	0.893	0.832	0.859	0.878	0.870
Numérica08	300	0.918	0.567	0.482	0.608	0.587	0.483	0.498
Numérica09	300	0.847	0.517	0.475	0.529	0.518	0.505	0.540
Numérica10	300	0.857	0.474	0.459	0.458	0.470	0.445	0.462
Combinada01	900	0.945	0.998	0.997	0.999	0.998	0.997	0.999
Combinada02**	900	0.450	0.702	0.672	0.754	0.736	0.699	0.724
Combinada03**	900	0.924	0.970	0.897	0.692*+	0.772*+	0.953	0.930
Combinada04	900	0.741	0.741	0.744	0.757	0.753	0.764	0.759
Combinada05	900	0.820	0.748	0.684	0.686	0.700	0.716	0.722

Tabla 21: Resultados del AUC con BD artificiales y clasificador K-NN.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Nominal01**	900	0.500	0.541	0.602	0.636	0.634	0.532	0.532
Nominal02**	900	0.507	0.405	0.420	0.471	0.471	0.404	0.404
Nominal03	400	0.499	0.577	0.565	0.563	0.556	0.590	0.590
Nominal04	400	0.496	0.499	0.522	0.518	0.535	0.497	0.497
Nominal05	400	0.500	0.526	0.518	0.523	0.523	0.537	0.537
Numérica01	800	0.950	1.000	1.000	1.000	1.000	1.000	1.000
Numérica02**	800	0.500	0.361*+	0.798	0.708*+	0.725*+	0.845	0.874
Numérica03**	800	0.507	0.555	0.494	0.569	0.579	0.508	0.520
Numérica04	800	0.500	0.496	0.456	0.440	0.459	0.467	0.452
Numérica05**	800	0.500	0.417	0.414	0.486	0.460	0.427	0.443
Numérica06	300	0.975	1.000	1.000	1.000	1.000	1.000	1.000
Numérica07**	300	0.500	0.569*+	0.893	0.832	0.859	0.878	0.870
Numérica08	300	0.640	0.567	0.482	0.608	0.587	0.483	0.498
Numérica09	300	0.500	0.517	0.475	0.529	0.518	0.505	0.540
Numérica10	300	0.491	0.474	0.459	0.458	0.470	0.445	0.462
Combinada01	900	0.950	0.998	0.997	0.999	0.998	0.997	0.999
Combinada02**	900	0.491	0.702	0.672	0.754	0.736	0.699	0.724
Combinada03**	900	0.642	0.970	0.897	0.692*+	0.772*+	0.953	0.930
Combinada04	900	0.716	0.741	0.744	0.757	0.753	0.764	0.759
Combinada05	900	0.674	0.748	0.684	0.686	0.700	0.716	0.722

Tabla 22: Resultados del AUC con BD artificiales y clasificador C4.5.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Nominal01**	900	0.500	0.541	0.602	0.636	0.634	0.532	0.532
Nominal02**	900	0.500	0.405	0.420	0.471	0.471	0.404	0.404
Nominal03	400	0.500	0.577	0.565	0.563	0.556	0.590	0.590
Nominal04	400	0.500	0.499	0.522	0.518	0.535	0.497	0.497
Nominal05	400	0.500	0.526	0.518	0.523	0.523	0.537	0.537
Numérica01	800	0.950	1.000	1.000	1.000	1.000	1.000	1.000
Numérica02**	800	0.500	0.361*+	0.798	0.708	0.725	0.845	0.874
Numérica03**	800	0.501	0.555	0.494	0.569	0.579	0.508	0.520
Numérica04	800	0.500	0.496	0.456	0.440	0.459	0.467	0.449
Numérica05**	800	0.500	0.417	0.414	0.486	0.460	0.427	0.443
Numérica06	300	0.975	1.000	1.000	1.000	1.000	1.000	1.000
Numérica07**	300	0.500	0.569*+	0.893	0.832	0.859	0.878	0.870
Numérica08	300	0.587	0.567	0.482	0.608	0.587	0.483	0.498
Numérica09	300	0.500	0.517	0.475	0.529	0.518	0.505	0.540
Numérica10**	300	0.491	0.474	0.459	0.458	0.470	0.445	0.462
Combinada01**	900	0.950	0.998	0.997	0.999	0.998	0.997	0.999
Combinada02**	900	0.492	0.702	0.672	0.754	0.736	0.699	0.724
Combinada03**	900	0.661	0.970	0.897	0.692*+	0.772	0.953	0.930
Combinada04	900	0.732	0.741	0.744	0.757	0.753*+	0.764	0.759
Combinada05	900	0.695	0.748	0.684	0.686	0.700	0.716	0.722

Tabla 23: Resultados del AUC con BD artificiales y clasificador PART.

Nombre	**%sob**	**Original**	**ROS**	**SMOTE**	**BSM1**	**BSM2**	**GIS-C**	**GIS-CJ**
Nominal01**	900	0.642	0.541	0.602	0.636	0.634	0.532	0.532
Nominal02**	900	0.457	0.405	0.420	0.471	0.471	0.404	0.404
Nominal03	400	0.566	0.577	0.565	0.563	0.556	0.590	0.590
Nominal04	400	0.506	0.499	0.522	0.518	0.535	0.497	0.497
Nominal05	400	0.525	0.526	0.518	0.523	0.523	0.537	0.537
Numérica01	800	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Numérica02**	800	0.240	0.361*+	0.798	0.708	0.725	0.845	0.874
Numérica03**	800	0.549	0.555	0.494	0.569	0.579	0.508	0.520
Numérica04	800	0.424	0.496	0.456	0.440	0.459	0.467	0.452
Numérica05**	800	0.380	0.417	0.414	0.486	0.460	0.427	0.443
Numérica06	300	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Numérica07**	300	0.582	0.569*+	0.893	0.832	0.859	0.878	0.870
Numérica08	300	0.526	0.567	0.482	0.608	0.587	0.483	0.498
Numérica09	300	0.475	0.517	0.475	0.529	0.518	0.505	0.540
Numérica10	300	0.460	0.474	0.459	0.458	0.470	0.445	0.462
Combinada01**	900	0.998	0.998	0.997	0.999	0.998	0.997	0.999
Combinada02**	900	0.648	0.702	0.672	0.754	0.736	0.699	0.724
Combinada03**	900	0.981	0.970	0.897	0.692*+	0.772*+	0.953	0.930
Combinada04	900	0.716	0.741	0.744	0.757	0.753	0.764	0.759
Combinada05	900	0.619	0.748	0.684	0.686	0.700	0.716	0.722

Tabla 24: Resultados del AUC con BD artificiales y clasificador *Backpropagation*.

APÉNDICE E

Resultados del recuerdo sobre bases de datos reales

Este apéndice contiene las tablas de resultados del recuerdo de los seis métodos de sobre-muestreo (ROS, SMOTE, BSM1, BSM2, GIS-G y GIS-GF) con las bases de datos de dominios reales. Son seis tablas, cada una de ellas corresponde a cada clasificador utilizado: *AdaBoost M1*, *Naive Bayes*, K-NN, C4.5, PART y *Backpropagation*.

Una marca ** delante del nombre de la base de datos indica que ANOVA mostró que los resultados en general son significativos, no se deben a la aleatoriedad introducida por los métodos de sobre-muestreo. Un signo * delante de cada valor del recuerdo indica que la diferencia de ese valor con el obtenido por GIS-G es significativa y un signo + que lo es con GIS-GF.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.350	0.465	0.460	0.455	0.450	0.465	0.470
Balloons**	100	0.729	0.731*+	0.797*+	0.806	0.817	0.823	0.823
Breast_Cancer	200	0.134	0.320*+	0.388	0.390	0.386	0.404	0.404
Bupa**	100	0.567	0.539*+	0.673*+	0.670*+	0.683	0.724	0.718
Car1**	700	0.629	0.920+	0.946	0.947	0.980	0.934	0.974
Car2**	700	0.339	0.991	0.943	0.951	0.900	0.970	0.940
Cardiovascular	400	0.072	0.193*+	0.510	0.537	0.543	0.470	0.458
Cinco**	300	0.410	0.200*+	0.525*+	0.555	0.635	0.615	0.630
Coil	200	0.217	0.172*+	0.244	0.236	0.264	0.242	0.244
Cpu**	1000	0.714	0.086*+	0.843*	0.586*+	0.629*+	0.771+	0.871*
Diabetes**	100	0.585	0.515*+	0.638*+	0.683	0.732	0.738	0.721
Ecoli**	600	0.516	0.724*+	0.816	0.800	0.816	0.818	0.829
Escalon**	300	0.085	0.365*+	0.805	0.785	0.840	0.795	0.825
Glass**	700	0.000	0.006*+	0.506	0.211*+	0.206*+	0.467	0.517
Haberman**	200	0.304	0.467*+	0.736	0.844*+	0.840*+	0.701	0.733
Hayes-roth**	200	1.000	1.000	1.000	0.987	0.980	1.000	0.997
Heart**	300	0.495	0.515*+	0.615	0.600	0.665	0.633	0.658
Hepatitis**	300	0.541	0.581+	0.600+	0.626	0.670	0.596+	0.648*
Imayuscula**	600	0.200	0.058*+	0.667*	0.392*+	0.450*+	0.567	0.633
Postoperative Patient	100	0.012	0.104	0.154	0.133	0.154	0.142	0.137
Raro**	100	0.738	0.383*+	0.746	0.679	0.762	0.704	0.700
Servo**	100	0.808	0.788*+	0.802+	0.804+	0.814+	0.814+	0.888*
Wine	100	0.988	0.981	0.990	0.990	0.990	0.988	0.988

Tabla 25: Resultados del recuerdo con BD reales y clasificador *AdaBoost M1*.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.440	0.465	0.460	0.455	0.450	0.465	0.470
Balloons**	100	0.737	0.731*+	0.797*+	0.806	0.817	0.823	0.823
Breast_Cancer	200	0.400	0.320*+	0.388	0.390	0.386	0.404	0.404
Bupa**	100	0.683	0.539*+	0.673*+	0.670*+	0.683*+	0.724	0.718
Car1**	700	0.603	0.920*+	0.946	0.947	0.974	0.934	0.974
Car2**	700	0.333	0.991	0.943	0.951	0.900*+	0.970	0.940
Cardiovascular	400	0.212	0.193*+	0.510	0.537	0.543	0.470	0.458
Cinco**	300	0.010	0.200*+	0.525*+	0.555*+	0.635	0.615	0.635
Coil	200	0.263	0.172*+	0.244	0.236	0.264	0.242	0.244
Cpu**	1000	0.857	0.086*+	0.843	0.586*+	0.629*+	0.771+	0.871*
Diabetes**	100	0.587	0.515*+	0.638*+	0.683*+	0.732	0.738	0.721
Ecoli**	600	0.847	0.724*+	0.816*	0.800*	0.816	0.818	0.829
Escalon**	300	0.000	0.365*+	0.805	0.785	0.840	0.795	0.825
Glass**	700	0.756	0.006	0.506	0.211	0.206	0.467	0.517
Haberman**	200	0.198	0.467*+	0.736*	0.844*+	0.840*+	0.701	0.733
Hayes-roth**	200	1.000	1.000	1.000	0.987	0.980	1.000	0.987
Heart**	300	0.664	0.515*+	0.615	0.600	0.665	0.633	0.658
Hepatitis**	300	0.669	0.581	0.600	0.626	0.670	0.596	0.648
Imayuscula**	600	0.000	0.058*+	0.667	0.392*+	0.450*+	0.567	0.633
Postoperative Patient	100	0.017	0.104	0.154	0.133	0.154	0.142	0.137
Raro**	100	0.208	0.383*+	0.746	0.679*+	0.762	0.704	0.700
Servo**	100	0.760	0.788*+	0.802	0.804	0.814	0.814	0.888
Wine	100	0.900	0.981	0.990	0.990	0.990	0.988	0.988

Tabla 26: Resultados del recuerdo con BD reales y clasificador *Naive Bayes*.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.380	0.465	0.460	0.455	0.450	0.465	0.470
Balloons**	100	0.774	0.731*+	0.797*+	0.806	0.817	0.823	0.823
Breast_Cancer	200	0.236	0.320*+	0.388*+	0.390*+	0.386*+	0.404	0.404
Bupa**	100	0.557	0.539*+	0.673*+	0.670*+	0.683*+	0.724	0.718
Car1**	700	0.784	0.920	0.946	0.947	0.980	0.934	0.974
Car2**	700	0.456	0.991+	0.943	0.951	0.900*+	0.970	0.940
Cardiovascular	400	0.113	0.193*+	0.510	0.537	0.543	0.470	0.458
Cinco**	300	0.200	0.200*+	0.525*+	0.555*+	0.635	0.615	0.630
Coil	200	0.114	0.172*+	0.244	0.236	0.264	0.242	0.244
Cpu**	1000	0.243	0.086*+	0.843	0.586*+	0.629*+	0.771	0.871
Diabetes**	100	0.557	0.515*+	0.638*+	0.683*+	0.732	0.738	0.721
Ecoli**	600	0.527	0.724*+	0.816+	0.800+	0.82	0.818	0.829
Escalon**	300	0.525	0.365*+	0.805	0.785	0.840	0.795	0.825
Glass**	700	0.033	0.006*+	0.506	0.211*+	0.206*+	0.467	0.517
Haberman**	200	0.257	0.467*+	0.736	0.844	0.840	0.701	0.733
Hayes-roth**	200	0.397	1.000	1.000	0.987	0.980	1.000	0.997
Heart**	300	0.630	0.515*+	0.615	0.600*+	0.665	0.633	0.658
Hepatitis**	300	0.585	0.581*+	0.600+	0.626	0.670	0.596+	0.648*
Imayuscula**	600	0.150	0.058*+	0.667	0.392*+	0.450*+	0.667	0.633
Postoperative Patient	100	0.017	0.104*+	0.154	0.133	0.154	0.142	0.137
Raro**	100	0.737	0.383*+	0.746	0.679*+	0.762	0.704	0.700
Servo**	100	0.774	0.788*+	0.802	0.804	0.814	0.814	0.888
Wine	100	0.998	0.981	0.990	0.990	0.990	0.988	0.988

Tabla 27: Resultados del recuerdo con BD reales y clasificador K-NN.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.385	0.465	0.460	0.455	0.450	0.465	0.470
Balloons**	100	0.774	0.731*+	0.797*+	0.806	0.817	0.823	0.823
Breast_Cancer	200	0.186	0.320*+	0.388	0.390	0.386	0.404	0.404
Bupa**	100	0.512	0.539*+	0.673*+	0.670*+	0.683*+	0.724	0.718
Cardiovascular	400	0.052	0.193*+	0.510	0.537	0.543	0.470	0.458
Coil	200	0.247	0.172*+	0.244	0.236	0.264	0.242	0.244
cpu	1000	0.871	0.086*+	0.843	0.586*+	0.629*+	0.771+	0.871*
Ecoli**	600	0.547	0.724*+	0.816	0.800	0.816	0.818	0.829
Glass**	700	0.033	0.006*+	0.506	0.211	0.206	0.467	0.517
Heart**	300	0.477	0.515*+	0.615+	0.600*+	0.665	0.633	0.658
Hepatitis**	300	0.478	0.581	0.600	0.626	0.670	0.596	0.648
Postoperative Patient	100	0.104	0.104	0.154	0.133	0.154	0.142	0.137
Car1**	700	0.916	0.920	0.946	0.947	0.980	0.934	0.974
Car2**	700	0.850	0.991	0.943	0.951	0.900*+	0.970	0.940
Cinco**	300	0.445	0.200*+	0.525*+	0.555*+	0.635	0.615	0.630
Diabetes**	100	0.557	0.515*+	0.638*+	0.683*+	0.732	0.738	0.721
Escalon**	300	0.105	0.365*+	0.805	0.785	0.840	0.795	0.825
Haberman**	200	0.417	0.467*+	0.736	0.844*+	0.840*+	0.701	0.733
Hayes-roth**	200	0.473	1.000	1.000	0.987	0.980	1.000	0.997
Imayuscula**	600	0.042	0.058*+	0.667	0.392*+	0.450*+	0.567	0.633
Raro**	100	0.692	0.383*+	0.746	0.679*+	0.762	0.704	0.700
Servo**	100	0.772	0.788+	0.802	0.804	0.814	0.814	0.888
Wine	100	0.883	0.981	0.990	0.990	0.990	0.988	0.988

Tabla 28: Resultados del recuerdo con BD reales y clasificador C4.5.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.370	0.465	0.460	0.455	0.450	0.465	0.470
Balloons**	100	0.740	0.731*+	0.797*+	0.806	0.817	0.823	0.823
Breast_Cancer	200	0.134	0.320*+	0.388	0.390	0.386	0.404	0.404
Bupa**	100	0.472	0.539*+	0.673*+	0.670*+	0.683*+	0.724	0.718
Car1**	700	0.890	0.920	0.946	0.947	0.980	0.934	0.974
Car2**	700	0.880	0.991	0.943	0.951	0.900*+	0.970	0.940
Cardiovascular	400	0.047	0.193*+	0.510	0.537	0.543	0.470	0.458
Cinco**	300	0.450	0.200*+	0.525*+	0.555*+	0.635	0.615	0.630
Coil	200	0.192	0.172*+	0.244	0.236	0.264	0.242	0.244
Cpu	1000	0.857	0.086*+	0.843	0.586*+	0.629*+	0.771	0.871
Diabetes**	100	0.615	0.515*+	0.638*+	0.683*+	0.732	0.738	0.721
Ecoli**	600	0.504	0.724*+	0.816	0.800	0.816	0.818	0.829
Escalon**	300	0.105	0.365*+	0.805	0.785	0.840	0.795	0.825
Glass**	700	0.033	0.006*+	0.506	0.211*+	0.206*+	0.467	0.517
Haberman**	200	0.284	0.467*+	0.736	0.844*+	0.840*+	0.701	0.733
Hayes-roth**	200	0.430	1.000	1.000	0.987	0.980	1.000	0.997
Heart**	300	0.500	0.515*+	0.615	0.600	0.665	0.633	0.658
Hepatitis**	300	0.533	0.581	0.600	0.626	0.670	0.596	0.648
Imayuscula**	600	0.075	0.058*+	0.667	0.392*+	0.450*+	0.567	0.633
Postoperative Patient	100	0.071	0.104	0.154	0.133	0.154	0.142	0.137
Raro**	100	0.692	0.383*+	0.746	0.679	0.762	0.704	0.700
Servo**	100	0.758	0.788	0.802	0.804	0.814	0.814	0.888
Wine	100	0.892	0.981	0.990	0.990	0.990	0.988	0.988

Tabla 29: Resultados del recuerdo con BD reales y clasificador PART.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.450	0.465	0.460	0.455	0.450	0.465	0.470
Balloons**	100	0.757	0.731*+	0.797*+	0.806	0.817	0.823	0.823
Breast_Cancer	200	0.310	0.320	0.388	0.390	0.386	0.404	0.404
Bupa**	100	0.521	0.539*+	0.673*+	0.670*+	0.683	0.724	0.718
Car1**	700	0.924	0.920	0.946	0.947	0.980	0.934	0.974
Car2**	700	0.900	0.991	0.943	0.951	0.900*+	0.970	0.940
Cardiovascular	400	0.073	0.193*+	0.510	0.537	0.543	0.470	0.458
Cinco**	300	0.155	0.200*+	0.525*+	0.555*+	0.635	0.615	0.630
Coil	200	0.142	0.172*+	0.244	0.236	0.264	0.242	0.244
Cpu	1000	0.000	0.086*+	0.843	0.586*+	0.629*+	0.771	0.871
Diabetes**	100	0.608	0.515*+	0.638*+	0.683*+	0.732	0.738	0.721
Ecoli**	600	0.711	0.724*+	0.816	0.800	0.816	0.818	0.829
Escalon**	300	0.145	0.365*+	0.805	0.785	0.840	0.795	0.825
Glass**	700	0.000	0.006*+	0.506	0.211*+	0.206*+	0.467	0.517
Haberman**	200	0.331	0.467	0.736	0.844	0.840	0.701	0.733
Hayes-roth**	200	1.000	1.000	1.000	0.987	0.980	1.000	0.997
Heart**	300	0.513	0.515	0.615	0.600	0.665	0.633	0.658
Hepatitis**	300	0.581	0.581	0.600	0.626	0.670	0.596	0.648
Imayuscula**	600	0.000	0.058*+	0.667	0.392*+	0.450*+	0.567	0.633
Postoperative Patient	100	0.133	0.104	0.154	0.133	0.154	0.142	0.137
Raro**	100	0.233	0.383*+	0.746	0.679	0.762	0.704	0.700
Servo**	100	0.798	0.788	0.802	0.804	0.814	0.814	0.888
Wine	100	0.990	0.981	0.990	0.990	0.990	0.988	0.988

Tabla 30: Resultados del recuerdo con BD reales y clasificador *Backpropagation*.

APÉNDICE F

Resultados *Precision* sobre bases de datos reales

Este apéndice contiene las tablas de resultados de la precisión de los seis métodos de sobre-muestreo (ROS, SMOTE, BSM1, BSM2, GIS-G y GIS-GF) con las bases de datos de dominios reales. Son seis tablas, cada una de ellas corresponde a cada clasificador utilizado: *AdaBoost M1*, *Naive Bayes*, K-NN, C4.5, PART y *Backpropagation*.

Una marca ** delante del nombre de la base de datos indica que ANOVA mostró que los resultados en general son significativos, no se deben a la aleatoriedad introducida por los métodos de sobre-muestreo. Un signo * delante de cada valor del recuerdo indica que la diferencia de ese valor con el obtenido por GIS-G es significativa y un signo + que lo es con GIS-GF.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.822	0.743*+	0.646	0.585*+	0.552*+	0.639	0.687
Balloons**	100	0.807	0.736	0.691	0.718*+	0.637	0.637	0.632
Breast_Cancer	200	0.282	0.289*+	0.302*+	0.315	0.289*+	0.347	0.342
Bupa**	100	0.719	0.692*+	0.602	0.592	0.603	0.579	0.585
Car1**	700	0.923	0.932+	0.918+	0.887+	0.812+	0.882+	0.772*
Car2**	700	0.915	0.989+	0.856*	0.846*	0.694*	0.929+	0.732*
Cardiovascular	400	0.234	0.279	0.181	0.189	0.177	0.212	0.200
Cinco**	300	0.657	0.312	0.245*+	0.259*+	0.284*+	0.341	0.342
Coil	200	0.283	0.192	0.176	0.225	0.206	0.196	0.192
Cpu**	1000	0.686	0.055*+	0.369+	0.285*+	0.252*+	0.345+	0.413*
Diabetes**	100	0.674	0.641	0.621	0.554	0.540	0.585	0.620
Ecoli**	600	0.772	0.709+*	0.616	0.683	0.653	0.606	0.595
Escalon**	300	0.136	0.467*	0.586+	0.552	0.508	0.508	0.487
Glass**	700	0.000	0.004+*	0.220	0.164	0.135	0.145	0.159
Haberman**	200	0.517	0.456*+	0.347	0.294*+	0.288*+	0.372	0.361
Hayes-roth**	200	1.000	1.000	1.000	0.919	0.764*+	1.000	0.971
Heart**	300	0.674	0.492	0.479	0.500	0.498	0.509	0.511
Hepatitis**	300	0.577	0.604	0.539	0.647+*	0.606	0.533	0.561
Imayuscula**	600	0.290	0.097*+	0.245	0.399*+	0.330*+	0.225	0.193
Postoperative Patient	100	0.021	0.137*+	0.150	0.155	0.165	0.174	0.162
Raro**	100	0.886	0.424	0.447	0.606	0.563	0.444	0.457
Servo**	100	0.864	0.820+	0.810+	0.818+	0.807+	0.829+	0.677*
Wine	100	0.948	0.966	0.969	0.968	0.968	0.966	0.969

Tabla 31: Resultados de la precisión con BD reales y clasificador *AdaBoost M1*.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.417	0.743	0.646	0.585*+	0.552*+	0.639	0.687
Balloons**	100	0.809	0.736	0.691	0.718	0.637	0.637	0.632
Breast_Cancer**	200	0.313	0.289*+	0.302*+	0.315*+	0.289*+	0.347	0.342
Bupa**	100	0.478	0.692	0.602	0.592	0.603	0.579	0.585
Car1**	700	0.739	0.932	0.918	0.887	0.812	0.882	0.772
Car2**	700	0.918	0.989	0.856*+	0.846*+	0.694*+	0.929	0.732
Cardiovascular	400	0.379	0.279	0.181	0.189	0.177	0.212	0.200
Cinco**	300	0.020	0.312	0.245*+	0.259*+	0.284*+	0.341	0.342
Coil	200	0.298	0.192	0.176	0.225	0.206	0.196	0.192
Cpu**	1000	0.177	0.055*+	0.369+	0.285*+	0.252*+	0.345	0.413
Diabetes**	100	0.658	0.641	0.621	0.554	0.540	0.585	0.620
Ecoli**	600	0.503	0.709+*	0.616	0.683	0.653	0.606	0.595
Escalon**	300	0.000	0.467	0.586	0.552	0.508	0.508	0.450
Glass**	700	0.102	0.004*+	0.220	0.164	0.135	0.145	0.159
Haberman**	200	0.543	0.456	0.347	0.294*+	0.288*+	0.372	0.361
Hayes-roth**	200	1.000	1.000	1.000	0.919	0.764*+	1.000	0.971
Heart**	300	0.497	0.492	0.479	0.500	0.498	0.509	0.511
Hepatitis**	300	0.660	0.604	0.539	0.647	0.606	0.533	0.561
Imayuscula**	600	0.000	0.097*+	0.245	0.399*+	0.330	0.225	0.193
Postoperative Patient	100	0.047	0.137	0.150	0.155	0.165	0.174	0.162
Raro**	100	0.464	0.424	0.447	0.606*+	0.563	0.444	0.457
Servo**	100	0.878	0.820+	0.810+	0.818+	0.807+	0.829+	0.677*
Wine	100	0.934	0.966	0.969	0.968	0.968	0.966	0.969

Tabla 32: Resultados de la precisión con BD reales y clasificador *Naive Bayes*.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.650	0.743*+	0.646	0.585*+	0.552*+	0.639	0.687
Balloons**	100	0.770	0.736*+	0.691	0.718*+	0.637	0.637	0.632
Breast_Cancer	200	0.315	0.289*+	0.302*+	0.315	0.289*+	0.347	0.342
Bupa**	100	0.548	0.692	0.602	0.592	0.603	0.579	0.585
Car1**	700	0.953	0.932	0.918	0.887	0.812	0.882	0.772
Car2**	700	0.889	0.989+	0.856*	0.846*	0.694*+	0.929	0.732
Cardiovascular	400	0.174	0.279	0.181	0.189	0.177*+	0.212	0.200
Cinco**	300	0.860	0.312	0.245*+	0.259*+	0.284*+	0.341	0.342
Coil	200	0.171	0.192	0.176	0.225	0.206	0.196	0.192
Cpu**	1000	0.136	0.055*+	0.369	0.285*+	0.252*+	0.345	0.413
Diabetes**	100	0.619	0.641	0.621	0.554	0.540	0.585	0.620
Ecoli**	600	0.662	0.709	0.616	0.683	0.653	0.606	0.595
Escalon**	300	0.803	0.467*	0.586	0.552	0.508	0.508	0.487
Glass**	700	0.061	0.004*+	0.220	0.164	0.135	0.145	0.159
Haberman**	200	0.389	0.456*+	0.347	0.294*+	0.288*+	0.372	0.361
Hayes-roth**	200	0.917	1.000	1.000	0.919	0.764*+	1.000	0.971
Heart**	300	0.528	0.492	0.479	0.500	0.498	0.509	0.511
Hepatitis**	300	0.572	0.604	0.539	0.647*+	0.606	0.533	0.561
Imayuscula**	600	0.268	0.097*+	0.245	0.399*+	0.330	0.225	0.193
Postoperative Patient	100	0.036	0.137	0.150	0.155	0.165	0.174	0.162
Raro**	100	0.785	0.424	0.447	0.606	0.563	0.444	0.457
Servo**	100	0.869	0.820	0.810	0.818	0.807	0.829	0.677
Wine	100	0.937	0.966	0.969	0.968	0.968	0.966	0.969

Tabla 33: Resultados de la precisión con BD reales y clasificador K-NN.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.640	0.743	0.646	0.585*+	0.552*+	0.639	0.687
Balloons**	100	0.737	0.736	0.691	0.718	0.637	0.637	0.632
Breast_Cancer	200	0.257	0.289*+	0.302	0.315	0.289*+	0.347	0.342
Bupa**	100	0.609	0.692	0.602	0.592	0.603	0.579	0.585
Cardiovascular	400	0.097	0.279	0.181	0.189	0.177	0.212	0.200
Coil	200	0.222	0.192	0.176	0.225	0.206	0.196	0.192
Cpu	1000	0.729	0.055*+	0.369	0.285	0.252	0.345	0.413
Ecoli**	600	0.724	0.709	0.616	0.683	0.653	0.606	0.595
Glass**	700	0.043	0.004*+	0.220	0.164	0.135	0.145	0.159
Heart**	300	0.511	0.492	0.479	0.500	0.498	0.509	0.511
Hepatitis**	300	0.526	0.604	0.539	0.647	0.606	0.533	0.561
Postoperative Patient	100	0.179	0.137	0.150	0.155	0.165	0.174	0.162
Car1**	700	0.879	0.932	0.918	0.887	0.812	0.882	0.772
Car2**	700	0.789	0.989	0.856	0.846	0.694	0.929	0.732
Cinco	300	0.580	0.312	0.245	0.259	0.284	0.341	0.342
Diabetes**	100	0.662	0.641	0.621	0.554	0.540	0.585	0.620
Escalon	300	0.123	0.467	0.586	0.552	0.508	0.508	0.487
Haberman**	200	0.420	0.456*+	0.347	0.294*+	0.288	0.372	0.361
Hayes-roth**	200	0.967	1.000	1.000	0.919	0.764*+	1.000	0.971
imayuscula**	600	0.037	0.097*+	0.245	0.399	0.330	0.225	0.193
raro**	100	0.860	0.424	0.447	0.606	0.563	0.444	0.457
Servo**	100	0.814	0.820	0.810	0.818	0.807	0.829	0.677
Wine	100	0.925	0.966	0.969	0.968	0.968	0.966	0.969

Tabla 34: Resultados de la precisión con BD reales y clasificador C4.5.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.748	0.743*+	0.646	0.585*+	0.552*+	0.639	0.687
Balloons**	100	0.742	0.736*+	0.691	0.718*+	0.637	0.637	0.632
Breast_Cancer	200	0.240	0.289	0.302	0.315	0.289	0.347	0.342
Bupa**	100	0.637	0.692	0.602	0.592	0.603	0.579	0.585
Car1**	700	0.880	0.932	0.918	0.887	0.812	0.882	0.772
Car2**	700	0.857	0.989+	0.856*+	0.846*+	0.694*+	0.929+	0.732*
Cardiovascular	400	0.076	0.279	0.181	0.189	0.177	0.212	0.200
Cinco**	300	0.594	0.312	0.245*+	0.259*+	0.284*+	0.341	0.342
Coil	200	0.185	0.192	0.176	0.225	0.206	0.196	0.192
Cpu	1000	0.729	0.055*+	0.369	0.285*+	0.252*+	0.345	0.413
Diabetes**	100	0.613	0.641	0.621	0.554	0.540	0.585	0.620
Ecoli**	600	0.662	0.709	0.616	0.683	0.653	0.606	0.595
Escalon**	300	0.088	0.467	0.586	0.552	0.508	0.508	0.487
Glass**	700	0.046	0.004*+	0.220	0.164	0.135	0.145	0.159
Haberman**	200	0.369	0.456*+	0.347	0.294*+	0.288*+	0.372	0.361
Hayes-roth**	200	0.741	1.000	1.000	0.919	0.764*+	1.000	0.971
Heart**	300	0.533	0.492	0.479	0.500	0.498	0.509	0.511
Hepatitis**	300	0.547	0.604*+	0.539	0.647*+	0.606	0.533	0.561
Imayuscula**	600	0.059	0.097*+	0.245	0.399	0.330+	0.225	0.193
Postoperative Patient	100	0.103	0.137	0.150	0.155	0.165	0.174	0.162
Raro**	100	0.846	0.424	0.447	0.606	0.563	0.444	0.457
Servo**	100	0.835	0.820	0.810	0.818	0.807	0.829+	0.677*
Wine	100	0.933	0.966	0.969	0.968	0.968	0.966	0.969

Tabla 35: Resultados de la precisión con BD reales y clasificador PART.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1**	100	0.730	0.743	0.646	0.585*+	0.552*+	0.639	0.687
Balloons**	100	0.748	0.736*+	0.691	0.718	0.637	0.637	0.632
Breast_Cancer	200	0.292	0.289*+	0.302	0.315	0.289*+	0.347	0.342
Bupa**	100	0.708	0.692	0.602	0.592	0.603	0.579	0.585
Car1**	700	0.930	0.932+	0.918+	0.887+	0.812	0.882+	0.772*
Car2**	700	0.987	0.989+	0.856*+	0.846*+	0.694*+	0.929+	0.732*
Cardiovascular	400	0.213	0.279	0.181	0.189	0.177	0.212	0.200
Cinco**	300	0.238	0.312	0.245*+	0.259*+	0.284*+	0.341	0.342
Coil	200	0.159	0.192	0.176	0.225	0.206	0.196	0.192
Cpu	1000	0.000	0.055*+	0.369	0.285*+	0.252*+	0.345	0.413
Diabetes**	100	0.639	0.641	0.621	0.554*+	0.540*+	0.585	0.620
Ecoli**	600	0.697	0.709	0.616	0.683	0.653	0.606	0.595
Escalon**	300	0.203	0.467	0.586	0.552	0.508	0.508	0.487
Glass**	700	0.000	0.004*+	0.220	0.164	0.135	0.145	0.159
Haberman**	200	0.536	0.456	0.347	0.294*+	0.288*+	0.372	0.361
Hayes-roth**	200	1.000	1.000	1.000	0.919	0.764	1.000	0.971
Heart**	300	0.519	0.492	0.479	0.500	0.498	0.509	0.511
Hepatitis**	300	0.631	0.604	0.539	0.647	0.606	0.533	0.561
Imayuscula**	600	0.000	0.097*+	0.245	0.399*+	0.330	0.225	0.193
Postoperative Patient	100	0.179	0.137	0.150	0.155	0.165	0.174	0.162
Raro**	100	0.292	0.424	0.447	0.606*+	0.563	0.444	0.457
Servo**	100	0.840	0.820	0.810	0.818	0.807	0.829	0.677
Wine	100	0.968	0.966	0.969	0.968	0.968	0.966	0.969

Tabla 36: Resultados de la precisión con BD reales y clasificador *Backpropagation*.

APÉNDICE G

Resultados de la Medida-F sobre bases de datos reales

Este apéndice contiene las tablas de resultados de la Medida-F de los seis métodos de sobre-muestreo (ROS, SMOTE, BSM1, BSM2, GIS-G y GIS-GF) con las bases de datos de dominios reales. Son seis tablas, cada una de ellas corresponde a cada clasificador utilizado: *AdaBoost M1*, *Naive Bayes*, K-NN, C4.5, PART y *Backpropagation*.

Una marca ** delante del nombre de la base de datos indica que ANOVA mostró que los resultados en general son significativos, no se deben a la aleatoriedad introducida por los métodos de sobre-muestreo. Un signo * delante de cada valor del recuerdo indica que la diferencia de ese valor con el obtenido por GIS-G es significativa y un signo + que lo es con GIS-GF.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.478	0.558	0.516	0.490*+	0.484*+	0.528	0.541
Balloons**	100	0.736	0.710	0.727	0.746	0.706	0.707	0.703
Breast_Cancer	200	0.170	0.295*+	0.329	0.335	0.321	0.357	0.354
Bupa**	100	0.624	0.589	0.627	0.618	0.629	0.636	0.637
Car1**	700	0.726	0.916	0.923	0.907	0.879	0.895	0.851
Car2**	700	0.465	0.990+	0.891*+	0.890*+	0.775*+	0.945+	0.814*
Cardiovascular	400	0.104	0.212*+	0.261	0.275	0.263	0.283	0.271
Cinco**	300	0.491	0.235*+	0.294*+	0.327	0.366	0.397	0.395
Coil	200	0.232	0.171	0.200	0.222	0.221	0.207	0.209
Cpu**	1000	0.695	0.064*+	0.480	0.355*+	0.336*+	0.438	0.522
Diabetes**	100	0.625	0.567*+	0.628	0.606	0.619	0.650	0.665
Ecoli**	600	0.580	0.694	0.686	0.716	0.708	0.681	0.679
Escalon**	300	0.089	0.373*+	0.651	0.618	0.612	0.584	0.587
Glass**	700	0.000	0.005*+	0.297	0.173*+	0.154*+	0.217	0.239
Haberman**	200	0.362	0.446	0.466	0.434	0.427*+	0.473	0.470
Hayes-roth**	200	1.000	1.000	1.000	0.946	0.844*+	1.000	0.982
Heart**	300	0.553	0.493*+	0.526	0.534	0.560	0.552	0.561
Hepatitis**	300	0.530	0.564	0.541	0.607+*	0.614+*	0.537	0.574
Imayuscula**	600	0.221	0.071*+	0.306	0.354	0.326	0.287	0.274
Postoperative Patient	100	0.015	0.109*+	0.143	0.128	0.145	0.143	0.135
Raro**	100	0.780	0.363*+	0.524	0.590	0.597	0.519	0.504
Servo**	100	0.819	0.787	0.790	0.796	0.793	0.806	0.756
Wine	100	0.965	0.971	0.978	0.977	0.977	0.975	0.976

Tabla 37: Resultados de la Medida-F con BD reales y clasificador *AdaBoost M1*.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.423	0.558	0.516	0.490*+	0.484*+	0.528	0.541
Balloons**	100	0.747	0.710	0.727	0.746	0.706	0.707	0.703
Breast_Cancer	200	0.355	0.295*+	0.329	0.335	0.321	0.357	0.354
Bupa**	100	0.583	0.589*+	0.627	0.618	0.629	0.636	0.637
Car1**	700	0.642	0.916	0.923	0.907	0.879	0.895	0.851
Car2**	700	0.466	0.990	0.891	0.890	0.775*+	0.945	0.814
Cardiovascular	400	0.256	0.212	0.261	0.275	0.263	0.283	0.271
Cinco**	300	0.013	0.235*+	0.294*+	0.327	0.366	0.397	0.395
Coil	200	0.353	0.171*+	0.200	0.222	0.221	0.207	0.209
Cpu**	1000	0.281	0.064*+	0.480	0.355+	0.336+	0.438	0.522
Diabetes**	100	0.618	0.567*+	0.628	0.606*+	0.619	0.650	0.665
Ecoli**	600	0.623	0.694	0.686	0.716	0.708	0.681	0.679
Escalon**	300	0.000	0.373*+	0.651	0.618	0.612	0.584	0.587
Glass**	700	0.180	0.005*+	0.297	0.173*+	0.154*+	0.217	0.239
Haberman**	200	0.275	0.446	0.466	0.434	0.427	0.473	0.470
Hayes-roth**	200	1.000	1.000	1.000	0.946	0.844	1.000	0.982
Heart**	300	0.596	0.493	0.526	0.534	0.560	0.552	0.561
Hepatitis**	300	0.676	0.564	0.541	0.607	0.614	0.537	0.574
Imayuscula**	600	0.000	0.071*+	0.306	0.354	0.326	0.287	0.274
Postoperative Patient	100	0.024	0.109*+	0.143	0.128	0.145	0.143	0.135
Raro**	100	0.267	0.363*+	0.524	0.590	0.597	0.519	0.504
Servo**	100	0.798	0.787	0.790	0.796	0.793	0.806	0.756
Wine	100	0.963	0.971	0.978	0.977	0.977	0.975	0.976

Tabla 38: Resultados de la Medida-F con BD reales y clasificador *Naive Bayes*.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.474	0.558	0.516	0.490	0.484	0.528	0.541
Balloons**	100	0.759	0.710	0.727	0.746	0.706	0.707	0.703
Breast_Cancer	200	0.256	0.295	0.329	0.335	0.321	0.357	0.354
Bupa**	100	0.548	0.589*+	0.627	0.618	0.629	0.636	0.637
Car1**	700	0.842	0.916	0.923	0.907	0.879	0.895	0.851
Car2**	700	0.581	0.990+	0.891*	0.890*+	0.775*+	0.945	0.814
Cardiovascular	400	0.134	0.212	0.261	0.275	0.263	0.283	0.271
Cinco**	300	0.895	0.235*+	0.294*+	0.327	0.366	0.397	0.395
Coil	200	0.129	0.171*+	0.200	0.222	0.221	0.207	0.209
Cpu**	1000	0.166	0.064*+	0.480	0.355*+	0.336*+	0.438	0.522
Diabetes**	100	0.585	0.567*+	0.628	0.606*+	0.619	0.650	0.665
Ecoli**	600	0.559	0.694	0.686	0.716*+	0.708+*	0.681	0.679
Escalon**	300	0.602	0.373*+	0.651	0.618	0.612	0.584	0.587
Glass**	700	0.041	0.005*+	0.297	0.173*+	0.154*+	0.217	0.239
Haberman**	200	0.299	0.446	0.466	0.434	0.427	0.473	0.470
Hayes-roth**	200	0.537	1.000	1.000	0.946*+	0.844*+	1.000	0.982
Heart**	300	0.561	0.493*+	0.526	0.534	0.560	0.552	0.561
Hepatitis**	300	0.545	0.564	0.541	0.607+*	0.614+*	0.537	0.574
Imayuscula**	600	0.187	0.071*+	0.306	0.354	0.326	0.287	0.274
Postoperative Patient	100	0.022	0.109	0.143	0.128	0.145	0.143	0.135
Raro**	100	0.737	0.363*+	0.524	0.590	0.597	0.519	0.504
Servo**	100	0.802	0.787	0.790	0.796	0.793	0.806	0.756
Wine	100	0.964	0.971	0.978	0.977	0.977	0.975	0.976

Tabla 39: Resultados de la Medida-F con BD reales y clasificador K-NN.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.468	0.558	0.516	0.490*+	0.484*+	0.528	0.541
Balloons**	100	0.742	0.710	0.727	0.746	0.706	0.707	0.703
Breast_Cancer	200	0.199	0.295*+	0.329	0.335	0.321	0.357	0.354
Bupa**	100	0.547	0.589*+	0.627	0.618	0.629	0.636	0.637
Cardiovascular	400	0.062	0.212	0.261	0.275	0.263	0.283	0.271
Coil	200	0.221	0.171	0.200	0.222	0.221	0.207	0.209
Cpu	1000	0.776	0.064*+	0.480	0.355*+	0.336*+	0.438	0.522
Ecoli**	600	0.589	0.694	0.686	0.716	0.708	0.681	0.679
Glass**	700	0.034	0.005*+	0.297	0.173	0.154*+	0.217	0.239
Heart**	300	0.474	0.493	0.526	0.534	0.560	0.552	0.561
Hepatitis**	300	0.467	0.564	0.541	0.607	0.614	0.537	0.574
Postoperative Patient	100	0.176	0.109	0.143	0.128	0.145	0.143	0.135
Car1**	700	0.886	0.916	0.923	0.907	0.879	0.895	0.851
Car2**	700	0.804	0.990	0.891*	0.890*	0.775*+	0.945	0.814
Cinco**	300	0.480	0.235*+	0.294*+	0.327	0.366	0.397	0.395
Diabetes**	100	0.595	0.567*+	0.628	0.606	0.619	0.650	0.665
Escalon**	300	0.103	0.373*+	0.651	0.618	0.612	0.584	0.587
Haberman**	200	0.407	0.446	0.466	0.434	0.427	0.473	0.470
Hayes-roth**	200	0.621	1.000	1.000	0.946	0.844*+	1.000	0.982
Imayuscula**	600	0.036	0.071*+	0.306	0.354	0.326	0.287	0.274
Raro**	100	0.741	0.363	0.524	0.590	0.597	0.519	0.504
Servo**	100	0.774	0.787	0.790	0.796	0.793	0.806	0.756
Wine	100	0.897	0.971	0.978	0.977	0.977	0.975	0.976

Tabla 40: Resultados de la Medida-F con BD reales y clasificador C4.5.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.476	0.558	0.516	0.490	0.484	0.528	0.541
Balloons**	100	0.721	0.710	0.727	0.746	0.706	0.707	0.703
Breast_Cancer	200	0.151	0.295*+	0.329	0.335	0.321	0.357	0.354
Bupa**	100	0.515	0.589	0.627	0.618	0.629	0.636	0.637
Car1**	700	0.870	0.916	0.923	0.907	0.879	0.895	0.851
Car2**	700	0.858	0.990	0.891*	0.890*	0.775*+	0.945	0.814
Cardiovascular	400	0.054	0.212	0.261	0.275	0.263	0.283	0.271
Cinco**	300	0.491	0.235	0.294	0.327	0.366	0.397	0.395
Coil	200	0.183	0.171	0.200	0.222	0.221	0.207	0.209
Cpu	1000	0.771	0.064*+	0.480	0.355*+	0.336*+	0.438	0.522
Diabetes**	100	0.600	0.567*+	0.628	0.606	0.619	0.650	0.665
Ecoli**	600	0.511	0.694	0.686	0.716	0.708	0.681	0.679
Escalon**	300	0.092	0.373*+	0.651	0.618	0.612	0.584	0.587
Glass**	700	0.036	0.005*+	0.297	0.173	0.154	0.217	0.239
Haberman**	200	0.302	0.446	0.466	0.434	0.427	0.473	0.470
Hayes-roth**	200	0.500	1.000	1.000	0.946	0.844*+	1.000	0.982
Heart**	300	0.493	0.493	0.526	0.534	0.560	0.552	0.561
Hepatitis**	300	0.504	0.564	0.541	0.607	0.614	0.537	0.574
Imayuscula**	600	0.062	0.071*+	0.306	0.354	0.326	0.287	0.274
Postoperative Patient	100	0.078	0.109	0.143	0.128	0.145	0.143	0.135
Raro**	100	0.733	0.363*+	0.524	0.590	0.597	0.519	0.504
Servo**	100	0.771	0.787	0.790	0.796	0.793	0.806	0.756
Wine	100	0.907	0.971	0.978	0.977	0.977	0.975	0.976

Tabla 41: Resultados de la Medida-F con BD reales y clasificador PART.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.574	0.558	0.516	0.490	0.484	0.528	0.541
Balloons**	100	0.738	0.710	0.727	0.746	0.706	0.707	0.703
Breast_Cancer	200	0.286	0.295	0.329	0.335	0.321	0.357	0.354
Bupa**	100	0.586	0.589	0.627	0.618	0.629	0.636	0.637
Car1**	700	0.918	0.916	0.923	0.907	0.879	0.895	0.851
Car2**	700	0.989	0.990	0.891	0.890	0.775	0.945	0.814
Cardiovascular	400	0.101	0.212	0.261	0.275	0.263	0.283	0.271
Cinco**	300	0.179	0.235	0.294	0.327	0.366	0.397	0.395
Coil	200	0.144	0.171	0.200	0.222	0.221	0.207	0.209
Cpu	1000	0.000	0.064	0.480	0.355	0.336	0.438	0.522
Diabetes**	100	0.620	0.567	0.628	0.606	0.619	0.650	0.665
Ecoli**	600	0.689	0.694	0.686	0.716	0.708	0.681	0.679
Escalon**	300	0.151	0.373	0.651	0.618	0.612	0.584	0.587
Glass**	700	0.000	0.005	0.297	0.173	0.154	0.217	0.239
Haberman**	200	0.394	0.446	0.466	0.434	0.427	0.473	0.470
Hayes-roth**	200	1.000	1.000	1.000	0.946	0.844	1.000	0.982
Heart**	300	0.501	0.493	0.526	0.534	0.560	0.552	0.561
Hepatitis**	300	0.585	0.564	0.541	0.607	0.614	0.537	0.574
Imayuscula**	600	0.000	0.071	0.306	0.354	0.326	0.287	0.274
Postoperative Patient	100	0.136	0.109	0.143	0.128	0.145	0.143	0.135
Raro**	100	0.234	0.363	0.524	0.590	0.597	0.519	0.504
Servo**	100	0.804	0.787	0.790	0.796	0.793	0.806	0.756
Wine	100	0.977	0.971	0.978	0.977	0.977	0.975	0.976

Tabla 42: Resultados de la Medida-F con BD reales y clasificador *Backpropagation*.

APÉNDICE H

Resultados del AUC sobre bases de datos reales

Este apéndice contiene las tablas de resultados del AUC de los seis métodos de sobre-muestreo (ROS, SMOTE, BSM1, BSM2, GIS-G y GIS-GF) con las bases de datos de dominios reales. Son seis tablas, cada una de ellas corresponde a cada clasificador utilizado: *AdaBoost M1*, *Naive Bayes*, K-NN, C4.5, PART y *Backpropagation*.

Una marca ** delante del nombre de la base de datos indica que ANOVA mostró que los resultados en general son significativos, no se deben a la aleatoriedad introducida por los métodos de sobre-muestreo. Un signo * delante de cada valor del recuerdo indica que la diferencia de ese valor con el obtenido por GIS-G es significativa y un signo + que lo es con GIS-GF.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.756	0.816	0.838	0.809	0.799	0.804	0.818
Balloons**	100	0.830	0.779	0.797	0.781	0.782	0.804	0.802
Breast_Cancer	200	0.559	0.596	0.604	0.606	0.585	0.636	0.632
Bupa**	100	0.774	0.735	0.728	0.722	0.732	0.726	0.728
Car1**	700	0.986	0.995	0.998	0.993	0.989	0.990	0.992
Car2**	700	0.969	0.999	0.993	0.994	0.983	0.998	0.987
Cardiovascular	400	0.588	0.582	0.532	0.524	0.514	0.549	0.558
Cinco**	300	0.796	0.506	0.497	0.559	0.569	0.584	0.571
Coil	200	0.687	0.576	0.552	0.582	0.575	0.570	0.566
Cpu**	1000	0.996	0.945	0.947	0.938	0.943	0.945	0.955
Diabetes**	100	0.810	0.782	0.795	0.774	0.774	0.804	0.802
Ecoli**	600	0.944	0.945	0.941	0.940	0.943	0.938	0.943
Escalon**	300	0.712	0.648*+	0.912	0.904	0.911	0.848	0.844
Glass**	700	0.793	0.660*+	0.788	0.735	0.733	0.730	0.739
Haberman**	200	0.652	0.656	0.677	0.662	0.669	0.691	0.696
Hayes-roth**	200	1.000	1.000	1.000	0.997	0.994	1.000	0.999
Heart**	300	0.840	0.797	0.808	0.802	0.794	0.814	0.816
Hepatitis**	300	0.801	0.810	0.762	0.824	0.817	0.769	0.783
Imayuscula**	600	0.876	0.363*+	0.577	0.612	0.633	0.621	0.610
Postoperative Patient	100	0.440	0.362	0.356	0.358	0.368	0.363	0.375
Raro**	100	0.891	0.614	0.650	0.718	0.728	0.657	0.649
Servo**	100	0.939	0.940	0.938	0.935	0.932	0.940	0.930
Wine	100	0.999	0.999	0.999	0.999	0.999	0.999	0.999

Tabla 43: Resultados del AUC con BD reales y clasificador *AdaBoost M1*.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.693	0.816	0.838	0.809	0.799	0.804	0.818
Balloons**	100	0.827	0.779	0.797	0.781	0.782	0.804	0.802
Breast_Cancer	200	0.619	0.596*+	0.604	0.606	0.585	0.636	0.632
Bupa**	100	0.638	0.735	0.728	0.722	0.732	0.726	0.728
Car1**	700	0.984	0.995	0.998	0.993	0.989	0.990	0.992
Car2**	700	0.966	0.999	0.993	0.994	0.983	0.998	0.987
Cardiovascular	400	0.643	0.582	0.532	0.524	0.514	0.549	0.558
Cinco**	300	0.732	0.506	0.497	0.559	0.569	0.584	0.571
Coil	200	0.672	0.576	0.552	0.582	0.575	0.570	0.566
Cpu**	1000	0.917	0.945	0.947	0.938	0.943	0.945	0.955
Diabetes**	100	0.809	0.782	0.795	0.774*+	0.774*+	0.804	0.802
Ecoli**	600	0.938	0.945	0.941	0.940	0.943	0.938	0.943
Escalon**	300	0.428	0.648*+	0.912	0.904	0.911	0.848	0.844
Glass**	700	0.730	0.660*+	0.788	0.735	0.733	0.730	0.739
Haberman**	200	0.658	0.656	0.677	0.662	0.669	0.691	0.696
Hayes-roth**	200	1.000	1.000	1.000	0.997	0.994	1.000	0.999
Heart**	300	0.851	0.797	0.808	0.802	0.794	0.814	0.816
Hepatitis**	300	0.886	0.810	0.762	0.824	0.817	0.769	0.783
Imayuscula**	600	0.632	0.363*+	0.577*+	0.612	0.633	0.621	0.610
Postoperative Patient	100	0.445	0.362	0.356	0.358	0.368	0.363	0.375
Raro**	100	0.731	0.614	0.650	0.718	0.728	0.657	0.649
Servo**	100	0.871	0.940	0.938	0.935	0.932	0.940	0.930
Wine	100	1.000	0.999	0.999	0.999	0.999	0.999	0.999

Tabla 44: Resultados del AUC con BD reales y clasificador *Naive Bayes*.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.715	0.816	0.838	0.809	0.799	0.804	0.818
Balloons**	100	0.798	0.779	0.797	0.781	0.782	0.804	0.802
Breast_Cancer	200	0.530	0.596*+	0.604	0.606	0.585*+	0.636	0.632
Bupa**	100	0.644	0.735	0.728	0.722	0.732	0.726	0.728
Car1**	700	0.996	0.995	0.998	0.993	0.989	0.990	0.992
Car2**	700	0.981	0.999	0.993	0.994	0.983	0.998	0.987
Cardiovascular	400	0.531	0.582	0.532	0.524	0.514	0.549	0.558
Cinco**	300	0.984	0.506	0.497*+	0.559	0.569	0.584	0.571
Coil	200	0.489	0.576	0.552	0.582	0.575	0.570	0.566
Cpu**	1000	0.889	0.945	0.947	0.938	0.943	0.945	0.955
Diabetes**	100	0.748	0.782	0.795	0.774*+	0.774*+	0.804	0.802
Ecoli**	600	0.885	0.945	0.941	0.940	0.943	0.938	0.943
Escalon**	300	0.916	0.648*+	0.912	0.904	0.911	0.848	0.844
Glass**	700	0.598	0.660*+	0.788	0.735	0.733	0.730	0.739
Haberman**	200	0.612	0.656	0.677	0.662	0.669	0.691	0.696
Hayes-roth**	200	0.997	1.000	1.000	0.997	0.994	1.000	0.999
Heart**	300	0.823	0.797	0.808	0.802	0.794	0.814	0.816
Hepatitis**	300	0.781	0.810	0.762	0.824	0.817	0.769	0.783
Imayuscula**	600	0.718	0.363*+	0.577*+	0.612	0.633	0.621	0.610
Postoperative Patient	100	0.251	0.362	0.356	0.358	0.368	0.363	0.375
Raro**	100	0.884	0.614	0.650	0.718	0.728	0.657	0.649
Servo**	100	0.939	0.940	0.938	0.935	0.932	0.940	0.930
Wine	100	0.994	0.999	0.999	0.999	0.999	0.999	0.999

Tabla 45: Resultados del AUC con BD reales y clasificador K-NN.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1**	100	0.690	0.816	0.838	0.809	0.799	0.804	0.818
Balloons**	100	0.747	0.779	0.797	0.781	0.782	0.804	0.802
Breast_Cancer	200	0.521	0.596	0.604	0.606	0.585	0.636	0.632
Bupa**	100	0.649	0.735	0.728	0.722	0.732	0.726	0.728
Cardiovascular	400	0.527	0.582	0.532	0.524	0.514	0.549	0.558
Coil	200	0.554	0.576	0.552	0.582	0.575	0.570	0.566
Cpu	1000	0.932	0.945	0.947	0.938	0.943	0.945	0.955
Ecoli**	600	0.785	0.945	0.941	0.940	0.943	0.938	0.943
Glass	700	0.589	0.660	0.788	0.735	0.733	0.730	0.739
Heart**	300	0.759	0.797	0.808	0.802	0.794	0.814	0.816
Hepatitis**	300	0.659	0.810	0.762	0.824	0.817	0.769	0.783
Postoperative Patient	100	0.412	0.362	0.356	0.358	0.368	0.363	0.375
Car1	700	0.966	0.995	0.998	0.993	0.989	0.990	0.992
Car2**	700	0.948	0.999	0.993	0.994	0.983	0.998	0.987
Cinco	300	0.706	0.506	0.497	0.559	0.569	0.584	0.571
Diabetes	100	0.756	0.782	0.795	0.774	0.774	0.804	0.802
Escalon**	300	0.540	0.648	0.912	0.904	0.911	0.848	0.844
Haberman**	200	0.614	0.656	0.677	0.662	0.669	0.691	0.696
Hayes-roth**	200	0.660	1.000	1.000	0.997	0.994	1.000	0.999
Imayuscula**	600	0.476	0.363	0.577	0.612	0.633	0.621	0.610
Raro**	100	0.853	0.614	0.650	0.718	0.728	0.657	0.649
Servo**	100	0.885	0.940	0.938	0.935	0.932	0.940	0.930
Wine	100	0.930	0.999	0.999	0.999	0.999	0.999	0.999

Tabla 46: Resultados del AUC con BD reales y clasificador C4.5.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.709	0.816	0.838	0.809	0.799	0.804	0.818
Balloons**	100	0.722	0.779	0.797	0.781	0.782	0.804	0.802
Breast_Cancer	200	0.567	0.596	0.604	0.606	0.585	0.636	0.632
Bupa**	100	0.667	0.735	0.728	0.722	0.732	0.726	0.728
Car1**	700	0.982	0.995	0.998	0.993	0.989	0.990	0.992
Car2**	700	0.980	0.999	0.993	0.994	0.983	0.998	0.987
Cardiovascular	400	0.526	0.582	0.532	0.524	0.514	0.549	0.558
Cinco**	300	0.711	0.506	0.497	0.559	0.569	0.584	0.571
Coil	200	0.498	0.576	0.552	0.582	0.575	0.570	0.566
Cpu	1000	0.947	0.945	0.947	0.938	0.943	0.945	0.955
Diabetes**	100	0.784	0.782	0.795	0.774	0.774	0.804	0.802
Ecoli**	600	0.798	0.945	0.941	0.940	0.943	0.938	0.943
Escalon**	300	0.539	0.648*+	0.912	0.904	0.911	0.848	0.844
Glass**	700	0.586	0.660*+	0.788	0.735	0.733	0.730	0.739
Haberman**	200	0.606	0.656	0.677	0.662	0.669	0.691	0.696
Hayes-roth**	200	0.715	1.000	1.000	0.997	0.994	1.000	0.999
Heart**	300	0.759	0.797	0.808	0.802	0.794	0.814	0.816
Hepatitis**	300	0.704	0.810	0.762	0.824	0.817	0.769	0.783
Imayuscula**	600	0.501	0.363*+	0.577	0.612	0.633	0.621	0.610
Postoperative Patient	100	0.427	0.362	0.356	0.358	0.368	0.363	0.375
Raro**	100	0.844	0.614	0.650	0.718	0.728	0.657	0.649
Servo**	100	0.903	0.940	0.938	0.935	0.932	0.940	0.930
Wine	100	0.939	0.999	0.999	0.999	0.999	0.999	0.999

Tabla 47: Resultados del AUC con BD reales y clasificador PART.

Nombre	%sob	Original	ROS	SMOTE	BSM1	BSM2	GIS-C	GIS-CJ
Abalone1	100	0.788	0.816	0.838	0.809	0.799	0.804	0.818
Balloons**	100	0.799	0.779	0.797	0.781	0.782	0.804	0.802
Breast_Cancer	200	0.603	0.596	0.604	0.606	0.585	0.636	0.632
Bupa**	100	0.739	0.735	0.728	0.722	0.732	0.726	0.728
Car1**	700	0.994	0.995	0.998	0.993	0.989	0.990	0.992
Car2**	700	0.999	0.999	0.993	0.994	0.983	0.998	0.987
Cardiovascular	400	0.561	0.582	0.532	0.524	0.514	0.549	0.558
Cinco**	300	0.487	0.506	0.497	0.559	0.569	0.584	0.571
Coil	200	0.569	0.576	0.552	0.582	0.575	0.570	0.566
Cpu	1000	0.938	0.945	0.947	0.938	0.943	0.945	0.955
Diabetes**	100	0.798	0.782	0.795	0.774	0.774	0.804	0.802
Ecoli**	600	0.946	0.945	0.941	0.940	0.943	0.938	0.943
Escalon**	300	0.569	0.648*+	0.912	0.904	0.911	0.848	0.844
Glass**	700	0.678	0.660	0.788	0.735	0.733	0.730	0.739
Haberman**	200	0.681	0.656	0.677	0.662	0.669	0.691	0.696
Hayes-roth**	200	1.000	1.000	1.000	0.997	0.994	1.000	0.999
Heart**	300	0.804	0.797	0.808	0.802	0.794	0.814	0.816
Hepatitis**	300	0.811	0.810	0.762	0.824	0.817	0.769	0.783
Imayuscula**	600	0.290	0.363*+	0.577*+	0.612	0.633	0.621	0.610
Postoperative Patient	100	0.367	0.362	0.356	0.358	0.368	0.363	0.375
Raro**	100	0.583	0.614	0.650	0.718	0.728	0.657	0.649
Servo**	100	0.943	0.940	0.938	0.935	0.932	0.940	0.930
Wine	100	0.999	0.999	0.999	0.999	0.999	0.999	0.999

Tabla 48: Resultados del AUC con BD reales y clasificador *Back-propagation*.