

METODOLOGÍA DE CLASIFICACIÓN DE DATOS DESBALANCEADOS BASADO EN MÉTODOS DE SUBMUESTREO.

Trabajo de grado para optar al título de Magíster en Ingeniería Eléctrica

Estudiante: Jhoan Keider Hoyos Osorio

Director: Genaro Daza Santacoloma

Co-director: Andrés Marino Álvarez Meza



**Universidad Tecnológica de Pereira
Facultad de Ingenierías - Programa de Ingeniería Eléctrica
Maestría en Ingeniería Eléctrica
Grupo de Investigación en Automática
Pereira, Risaralda, Colombia
2019**

Contenido

I	Preliminares	1
1.	Introducción	2
1.1.	Motivación	2
1.2.	Problema	3
2.	Objetivos	6
2.1.	Objetivo general	6
2.2.	Objetivos específicos	6
II	Marco Teórico	7
3.	Clasificación de datos desbalanceados	8
3.1.	Clasificación	8
3.2.	Desbalance de clases	9
3.3.	Métodos para abordar el desbalance de clases	10
3.3.1.	Técnicas de remuestreo	10
3.3.2.	Aprendizaje costo-sensitivo	11
3.3.3.	Métodos de ensamble	12
3.4.	Medidas de rendimiento de clasificación para el desbalance de clases	12

4. Métodos de remuestreo	15
4.1. SMOTE	15
4.2. Submuestreo basado en agrupamiento	16
4.2.1. Variante 1	16
4.2.2. Variante 2	17
4.3. Discusión y consideraciones de los métodos de remuestreo	18
5. Métodos de ensamble	19
5.1. Bagging	20
5.2. AdaBoost	20
6. Metodología de clasificación de datos desbalanceados a partir de submuestreo basado en agrupamiento	21
7. Nueva propuesta de submuestreo basado en el principio de información relevante (RIS)	23
7.1. Submuestreo basado en el principio de información relevante	24
7.2. Parámetros libres	27
8. CRIS: Combinación del RIS, la técnica de CBUS y el método de ensamble Bagging	28
III Marco Experimental	30
9. Esquema de trabajo	31
9.1. Clasificación de bases de datos desbalanceadas	31
9.1.1. Descripción de las bases de datos	32
9.1.2. Descripción de las pruebas de submuestreo	34

9.1.3. Sintonización de parámetros libres	34
9.1.4. Descripción de las pruebas de clasificación	35
9.1.5. Pruebas sobre la base de datos de FCD	36
9.1.6. Métodos de comparación	37
10.Resultados de las pruebas sobras bases de datos de pequeña escala	38
10.1. Resultados base de datos sintética	38
10.2. Resultados de las pruebas sobre las bases de datos del repositorio KEEL . .	40
11.Resultados de las pruebas sobre las bases de datos de gran escala	45
11.1. Resultados bases de datos de cáncer y proteínas	45
11.2. Resultados base de datos de FCD	47
IV Conclusiones	49
12.Conclusiones	50
A. Teoría de la información	52

Índice de figuras

4.1. Ejemplo del CBUS: Las muestras azules representan la clase mayoritaria, las rojas representan a las de la clase minoritaria, y las equis representan las observaciones seleccionadas por el método de muestreo.	17
6.1. Diagrama metodología de clasificación de datos desbalanceados a partir del submuestreo basado en agrupamiento y la técnica <i>Bagging</i>	22
7.1. Ejemplo ilustrativo. Se muestra el funcionamiento del algoritmo en la (a) 1 ^{ra} , (b) 10 ^{ava} , (c) 50 ^{ava} y (d) 655 ^{ava} iteración.	26
8.1. Diagrama metodología de submuestreo y clasificación de datos desbalanceados a partir de la técnica CRIS.	29
9.1. (a) Número de iteraciones hasta la convergencia, y (b) AUC alcanzadas por diferentes valores de λ	36
10.1. Resultados datos sintéticos	39
10.2. Boxplot de los resultados de clasificación	43
11.1. Resultados de clasificación para la base de datos de cáncer.	46
11.2. Resultados de clasificación para la base de datos de proteínas.	46
11.3. Curvas ROC de los diferentes métodos en estudio.	48

Índice de tablas

3.1. Matriz de confusión	13
9.1. Información de las bases de datos de pequeña escala	33
10.1. Rendimiento de clasificación de las diferentes clasificadores para las bases de datos de prueba	41
10.2. Rendimiento de clasificación de diferente número de clasificadores en el método CRIS.	42
10.3. Rendimiento de clasificación de las diferentes técnicas para las bases de datos de prueba	44
11.1. Comparación de resultados	47

Agradecimientos

Quiero agradecer a Dios por darme la oportunidad de culminar este trabajo. A mi familia por el apoyo y las oportunidas brindadas para formarme como un ingeniero físico y ahora como *Magíster*. A mi esposa Karen Velásquez quien ha estado a mi lado en esta etapa tan importante.

A mi tutor Genaro Daza Santacoloma y a mi codirector Andrés Marino Álvarez meza. Gracias por el conocimiento y la motivación, sus aportes han sido fundamentales para el desarrollo de este trabajo.

De igual manera manifiesto mi gratitud al grupo de investigación en Automática y a su director Álvaro Á. Orozco Gutiérrez, quienes pusieron a mi disposición las herramientas necesarias para el desarrollar este trabajo.

Esta investigación es desarrollado bajo el proyecto *Desarrollo de un sistema de identificación de estructuras nerviosas en imágenes de ultrasonido para la asistencia de bloqueo de nervios periféricos. Aplicación al tratamiento de dolor agudo traumático y prevención del dolor neuropático crónico* financiado por COLCIENCIAS con código 1110 – 744 – 55958. También, agradezco al programa de maestría en ingeniería eléctrica de la universidad tecnológica de Pereira y a la convocatoria interna de posgrados E6-18-5 financiada por la Vicerrectoría de Investigaciones, Innovación y Extensión.

Resumen

En este trabajo se presenta la construcción metodológica para la clasificación de datos desbalanceados, a partir del análisis comparativo entre técnicas de submuestreo, y tiene como aporte fundamental el desarrollo de una nueva estrategia de submuestreo y la clara identificación de las condiciones de aplicación de cada una de las técnicas. En particular, se consideran las técnicas de submuestreo basado en agrupamiento, un nuevo método de submuestreo basado en teoría de la información y una adaptación de los métodos propuesto para desarrollar un ensamble de clasificadores. Las pruebas de desempeño se orientan a la precisión del sistema en la etapa de clasificación y a la capacidad de cada método para seleccionar las muestras más representativas. Se realizan pruebas sobre 44 bases de datos desbalanceadas de pequeña escala del repositorio de datos KEEL, y tres bases de datos de gran escala orientadas a la predicción de cáncer de mama y de homología de proteínas y la detección automática de displasias corticales. Los resultados obtenidos reflejan que el submuestreo basado en teoría de la información es el método de submuestreo que mejor preserva la estructura de la clase mayoritaria, reduciendo la pérdida de información en el proceso de eliminación de muestras. Además, este método presenta una mejora sustancial cuando es adaptado para generar la combinación de diferentes clasificadores aumentando notablemente la capacidad del sistema para generalizar el comportamiento de ambas clases lo cual se puede evidenciar en los resultados de clasificación.

Abstract

This work presents a methodological construction for imbalanced data classification, which is founded on comparative analysis between subsampling techniques. Its main contribution is the development of a new subsampling strategy and a clearer identification of the conditions a specific method can be correctly applied. In particular, a clustering-based and a theory information-based undersampling techniques and an adaptation of the proposed methods to develop an ensemble are considered. Performance tests have their basis on classification accuracy and the capacity of each method to select the most representative samples. Experimental results are derived from 44 small-scale imbalanced datasets from KEEL data repository, and three large scale datasets oriented to breast cancer and protein homology prediction and the automatic detection of cortical dysplasias. Results show that information theory-based subsampling preserves the internal structure of data, and it reduces the loss of information by selecting the most informative samples. Also, this method presents a substantial improvement when it is adapted to generate a classifiers ensemble, increasing the system generalization capability, which can be evidenced in the classification results.

Parte I

Preliminares

Capítulo 1

Introducción

1.1. Motivación

En la tarea de reconocimiento de patrones y aprendizaje de máquina es muy común encontrar conjuntos de datos en los cuales alguna de las clases de interés tiene más muestras que la otra. Este fenómeno puede deberse a la poca frecuencia con la que ocurre algún evento, a la poca disponibilidad de muestras con un cierto atributo o a la dificultad para la recolección de los datos de esa clase. Estos patrones inusuales son generalmente difíciles de detectar por los algoritmos tradicionales, debido al desbalance entre la cantidad de información entre las clases [1]. En términos de clasificación, cualquier conjunto de datos que exhiba una distribución desigual entre sus clases puede considerarse desbalanceado. En un conjunto de datos desbalanceado, la clase mayoritaria tiene un gran porcentaje para todas las muestras, mientras que la clase minoritaria sólo ocupan una pequeña parte del total de las observaciones. Un caso específico de este tipo de fenómenos es la detección de transacciones fraudulentas; en esta aplicación una transacción fraudulenta podría aparecer entre mil transacciones válidas, por lo tanto al construir la base de datos se tendría un conjunto con un gran desequilibrio entre clases. En este ejemplo, identificar las transacciones fraudulentas es la prioridad, sin embargo, debido a la desproporción de información, los métodos de clasificación comunes tienen más probabilidad de clasificar nuevas observaciones como pertenecientes a la clase con más muestras. Esto representa un gran problema ya que el no identificar una transacción inválida podría ocasionar grandes pérdidas para los usuarios.

El problema del desbalance de clases tiene una importancia crucial, ya que se encuentra en un gran número de dominios de gran importancia ambiental, comercial, biomédico entre otros, y se ha demostrado, en ciertos casos, causar un cuello de botella significativo en el rendimiento alcanzable por los métodos de aprendizaje de máquina estándar que asumen

una distribución equilibrada de clases. Aplicaciones tales como la identificación de defectos de software, predicción de desastres naturales, reconocimiento automático de expresiones de genes de cáncer, detección de focos epilépticos en imágenes de resonancias magnéticas, transacciones de tarjetas de crédito fraudulentas, fraude de telecomunicaciones entre otros [2]. Por lo tanto, hay una gran motivación por parte de la industria y de la academia para desarrollar técnicas y metodologías para contrarrestar los efectos del desbalance de clases.

Desde el punto de vista académico, a pesar del crecimiento de la investigación en esta área durante los últimos 10 años, aún hay preguntas que no logran responderse, cómo lo son ¿Qué tipo de suposiciones harán que los algoritmos de aprendizaje desbalanceados funcionen mejor en comparación con el aprendizaje de las distribuciones originales? ¿Hasta qué punto debería balancearse el conjunto de datos original? entre muchos otros interrogantes.

Estas razones justifican la necesidad de desarrollar una metodología para la clasificación de datos desbalanceados que permita abordar aplicaciones de este tipo, brindando altas tasas de clasificación de la clase minoritaria, mediante técnicas avanzadas de re-muestreo, que conserven la mayor información de los datos, sin modificar la distribución original de los datos, evitando que los algoritmos estándar de clasificación se sesguen a la clase mayoritaria.

1.2. Problema

Las técnicas de clasificación suelen suponer que las muestras de entrenamiento están distribuidas uniformemente entre diferentes clases, por lo que generalmente un clasificador funciona bien cuando la técnica de clasificación se aplica a un conjunto de datos distribuido uniformemente entre sus diferentes clases.

El principal problema que ocasiona una base de datos desbalanceada es la capacidad de dichos datos para comprometer significativamente el rendimiento de la mayoría de los algoritmos de aprendizaje estándar. Debido a que la mayoría de los algoritmos estándar suponen o esperan distribuciones equilibradas de clases o costos iguales de clasificación errónea, cuando se presentan con conjuntos complejos de datos desbalanceados estos algoritmos no representan adecuadamente las características distributivas de los datos y por consiguiente, proporcionan precisiones desfavorables a través de las clases [3]. En realidad, se puede evidenciar que los clasificadores tienden a proporcionar un grado severamente desbalanceado de precisión, es decir un clasificador usualmente tenderá a predecir que cualquier muestra nueva, pertenecerá a la clase mayoritaria e ignorará completamente la clase minoritaria, lo cual reduce por completo el rendimiento de cualquier clasificador [4]. Este problema es aún mayor si se considera que la clase minoritaria en la mayoría de los casos es la de más relevancia y clasificar erróneamente patrones de esta naturaleza puede resultar en costos elevados.

Debido a la gran importancia de este problema, se han propuesto varios métodos en la literatura para tratar de solucionarlo. Estos métodos intentan dar una solución modificando los algoritmos de aprendizaje o el conjunto de datos. Al nivel de algoritmos se denominan como clasificadores costo-sensitivos [5]. Un clasificador costo-sensitivo aprende más características de las muestras de la clase minoritaria en comparación con la clase mayoritaria. Esto se hace estableciendo un alto costo a la clasificación errónea de una muestra de la clase minoritaria. No obstante, estos costos son a menudo desconocidos, y un clasificador de este tipo puede resultar en sobre-entrenamiento [6]. A nivel de datos, hay varios enfoques para tratar el problema de desequilibrio, que se dividen en dos grupos: enfoques basados en ensambles y enfoques basados en muestreo. El ensamble de clasificadores es una técnica útil para mejorar la precisión de la predicción al combinar varios clasificadores base [7]. Cada clasificador se entrena por separado, y la decisión final se toma por mayoría de votos. Estos métodos tienen la virtud de disminuir el riesgo de sobre-entrenamiento, sin embargo, los métodos de ensamble suelen ser computacionalmente costosos. Los enfoques de muestreo son técnicas de preprocesamiento, donde la distribución de los datos se reequilibra para reducir el efecto del desbalance en el proceso de aprendizaje [8]. Los métodos de muestreo se dividen en enfoques de sobremuestreo y submuestreo. Los métodos de sobremuestreo aumentan el número de muestras de la clase minoritaria, sin embargo, estas técnicas pueden aumentar la probabilidad de sobre-ajuste en el proceso de construcción del modelo [9]. Por el contrario, los enfoques de submuestreo reducen el número de muestras de la clase mayoritaria. El método más simple es conocido como muestreo aleatorio (RUS) [10], el cual elimina aleatoriamente muestras de la clase mayoritaria, generando un nuevo conjunto de entrenamiento. Naturalmente, el principal problema de este tipo de métodos es que datos útiles pertenecientes a la clase mayoritaria son eliminados, ya que no se considera la estructura subyacente de las observaciones. Recientemente, se han desarrollado estrategias de submuestreo que buscan preservar la estructura general de los datos y reducir la pérdida de información. Algunas de estas crean reglas a través de distancias para eliminar las muestras ya sea redundantes o irrelevantes. Una de estas son las técnicas de submuestreo basada en agrupamiento, las cuales realizan el submuestreo mediante la partición del conjunto de datos en varios conglomerados que codifican la estructura global de la clase mayoritaria. Sin embargo, estos métodos siguen siendo insuficientes para preservar la estructura interna ya que estos métodos codifican la estructura de la clase mayoritaria según distancias o minimizando funciones de costo que solo consideran estadísticas de segundo orden (medias y varianzas), por lo que no se explota al máximo la información de la clase mayoritaria.

Basados en estas problemáticas se plantea la siguiente pregunta de investigación: ¿cómo desarrollar una metodología de aprendizaje de máquina para la clasificación de conjuntos de datos desbalanceados, basada en técnicas de muestreo, que permita identificar estructuras relevantes de datos, seleccionar las muestras más informativas y que evite el sobre ajuste de entrenamiento en términos de medidas de sensibilidad y especificidad?

Por tanto, en este trabajo se propone el análisis y consideraciones de diferentes metodologías comúnmente aplicadas para la clasificación binaria de bases de datos desbalanceadas, lo cual ayuda a la generación de un esquema metodológico, con aplicabilidad a múltiples propósitos relacionados con el reconocimiento automático de patrones inusuales. Dicha metodología se basa principalmente en métodos de submuestreo y técnicas de ensamble de clasificadores que permitan disminuir los efectos relacionados a las características de los datos de interés en este estudio.

Este documento aborda en una primera etapa (Capítulo 3) la definición formal de la clasificación de datos desbalanceados; de igual manera, se exponen los principales enfoques para dar solución al problema de clasificación de datos desbalanceados. Posteriormente, se expone en el Capítulo 4 algunos algoritmos de submuestreo avanzados, los cuales han sido ampliamente utilizados para modificar el tamaño de las bases de datos, de forma precisa se señala el algoritmo SMOTE - *Synthetic Minority Oversampling Technique*, y el submuestreo basado en agrupamiento o conglomerados. Después, en el Capítulo 5 se exponen las principales técnicas de ensamble de clasificadores en combinación con diversos algoritmos de muestreo (*AdaBoosting* y *Bagging*).

En el Capítulo 7, se propone una metodología innovadora de submuestreo, la cual está basada en un principio de información relevante para preservar las estructuras más representativas de la clase con más datos a partir de la minimización de una función de costo en el marco del aprendizaje por teoría de la información. Es así como los Capítulos del 3 al 7 estructuran el marco teórico del documento.

Posteriormente, la Parte III del documento muestra las pruebas y los resultados experimentales, generados a partir de la comparación de los diferentes métodos de submuestreo. Estas comparaciones se llevan a cabo sobre bases de datos desbalanceadas del mundo real de diversas aplicaciones. La Parte IV presenta la discusión y conclusiones del trabajo y finalmente en la Parte V se tienen los apéndices, que son documentos claves para profundizar o comprender mejor algunas de las técnicas enunciadas.

Capítulo 2

Objetivos

2.1. Objetivo general

Desarrollar un sistema automático de aprendizaje de máquina para la clasificación de conjuntos de datos desbalanceados, basada en técnicas de muestreo, que permita identificar estructuras relevantes de datos y que evite el sobre ajuste de entrenamiento en términos de medidas de sensibilidad y especificidad.

2.2. Objetivos específicos

1. Desarrollar una metodología de sub-muestreo basada en técnicas de agrupamiento, con el fin de codificar la estructura global de datos en tareas de clasificación desbalanceadas.
2. Desarrollar una metodología de sub-muestreo basada en técnicas de información relevante, con el fin de representar la distribución y estructura local de la clase mayoritaria, para reducir la pérdida de información en la tarea de muestreo.
3. Formular una metodología discriminante de sub-muestreo para problemas de clasificación desbalanceadas, a partir del uso de métodos de ensamble, para preservar la estructura global y local de la clase mayoritaria y favorecer la escalabilidad del sistema de aprendizaje.

Parte II

Marco Teórico

Capítulo 3

Clasificación de datos desbalanceados

3.1. Clasificación

El aprendizaje de máquina es el sub-campo de las ciencias de la computación y una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. De manera más concreta, se trata de crear metodologías capaces de generalizar comportamientos y reconocer patrones a partir de una información suministrada en forma de ejemplos. Una rama del aprendizaje de máquina es el aprendizaje supervisado, el cual busca encontrar una función a partir de un conjunto de datos de entrenamiento, los cuales consisten en pares de objetos: una componente del par son los datos de entrada y el otro, los resultados deseados. El objetivo del aprendizaje supervisado es crear una función que pueda predecir la salida correspondiente a cualquier entrada válida después de haber sido sometido a una serie de datos ejemplos (datos de entrenamiento), es decir, el sistema generaliza con base en los datos que no ha visto. Este tipo de aprendizaje soluciona principalmente tareas de regresión donde la salida son valores numéricos, y de clasificación donde la salida es una etiqueta de clase.

En la teoría del aprendizaje de máquina, la clasificación es el problema de establecer una regla para identificar a cual clase corresponde una nueva observación, basados en el conjunto de datos de entrenamiento cuyas categorías son conocidas. Los algoritmos de clasificación permiten abstraer la información, llevándola a una representación adecuada para la toma de decisiones.

Algunos ejemplos de sistemas de clasificación son etiquetado automático de piezas o producto industrial como correcto o defectuoso [11], sistemas de seguridad para identificar si una persona tiene acceso o no a cierto lugar [12], detección de tumores en rayos-X [13], clasificación de pacientes como enfermos o no [14], entre otros.

3.2. Desbalance de clases

En la tarea de reconocimiento de patrones y aprendizaje de máquina existen diversos patrones inusuales, los cuales son difíciles de detectar debido a que estos eventos ocurren con mucha menos frecuencia que los que suceden comúnmente [1]. Fenómenos tales como desastres naturales, expresiones de genes de cáncer, transacciones fraudulentas, la ocurrencia de piezas defectuosas en un proceso industrial, son fenómenos que ocurren con menos regularidad que sus respectivos casos contrarios. Esta situación induce bases de datos que naturalmente tendrán más observaciones de los casos “normales” que de estos patrones inusuales.

En términos de clasificación, cualquier conjunto de datos que exhiba una distribución desigual entre sus clases puede considerarse desbalanceado. Este fenómeno se produce cuando hay más muestras en una clase que la otra clase en un conjunto de datos de entrenamiento. En un conjunto de datos desbalanceado, la clase mayoritaria tiene un gran porcentaje de todas las muestras, mientras que las instancias de la clase minoritaria sólo ocupan una pequeña parte de las observaciones.

Por lo general, sin consideración del problema del desbalance de clases, un algoritmo de clasificación tenderá a predecir que las muestras desconocidas pertenecen a la clase mayoritaria e ignoran completamente la clase minoritaria. Sin embargo, en muchas de las aplicaciones, la clase minoritaria es de vital importancia. Un ejemplo clásico es la clasificación de pacientes con cáncer o sanos a partir de imágenes de mamografías. Basados en la experiencia, el número de pacientes sanos supera notablemente el número de pacientes con cáncer. En consecuencia, en muchos algoritmos estándar de aprendizaje, se encuentran clasificadores que proveen un grado severamente desbalanceado de acierto, con efectividades de casi el 100 % para la clase mayoritaria pero efectividades entre 0 % y 10 %, aún cuando el hecho de saber si el paciente tiene cáncer es de tanta importancia.

Muchos otros conjuntos de datos en aplicaciones reales implican bases de datos con estas características, con aplicaciones tales como la identificación de defectos de software, predicción de desastres naturales, detección de focos epilépticos en imágenes de resonancias magnéticas, transacciones de tarjetas de crédito fraudulentas y fraude de telecomunicaciones [2]. Como se mencionó anteriormente, los algoritmos de aprendizaje de máquina con los que se suele abordar aplicaciones de este tipo, tienden a ignorar la clase minoritaria, que en estos casos es la de mayor relevancia y clasificar erróneamente eventos de esta naturaleza puede resultar en costos elevados. Un caso específico es la detección de fraude financiero, una transacción no válida podría surgir entre cientos de miles de registros de transacciones, pero el no identificar una transacción fraudulenta causaría enormes pérdidas.

3.3. Métodos para abordar el desbalance de clases

Dada la gran importancia del problema del desbalance de clases en diversos campos de aplicación, distintos algoritmos y métodos han sido propuestos en la última década para abordar el problema de clasificación de datos desbalanceados. En este dominio, se requiere un clasificador que proporcione una alta precisión para la clase minoritaria pero sin poner en peligro la precisión de la clase mayoritaria. En tal sentido, se han propuesto tres estrategias básicas: (I) *métodos de remuestreo*, los cuales son técnicas de preproceso que intentan equilibrar las distribuciones al considerar las proporciones representativas de los ejemplos de clase en la distribución, (II) *métodos de aprendizaje costo-sensitivos*, los cuales consideran los costos asociados con la clasificación errónea de las muestras [3] y (III) *métodos de ensamble* que consisten en la combinación de dos o más clasificadores.

A continuación se describen algunas de las técnicas contempladas en estos 3 enfoques:

3.3.1. Técnicas de remuestreo

Las técnicas de remuestreo se utilizan para equilibrar el espacio muestral para un conjunto de datos desequilibrado con el fin de aliviar el efecto de la distribución sesgada de clase en el proceso de aprendizaje. Los métodos de remuestreo son más versátiles porque son independientes del clasificador seleccionado [7]. Las técnicas de remuestreo se dividen en tres grupos dependiendo del método utilizado para equilibrar la distribución de clases:

- Métodos de sobremuestreo: Consisten en la creación de nuevas muestras de clase minoritaria. Dos métodos ampliamente utilizados para crear las muestras minoritarias sintéticas son duplicando al azar las muestras minoritarias y SMOTE (*Synthetic Minority Oversampling Technique*) [15], [16].

Los métodos de sobremuestreo poseen una gran dificultad, y es que añadir muestras a la clase minoritaria para balancear la base de datos, crea datos que no se puede asegurar que provengan de la distribución original, generando ruido para los clasificadores lo cual podría resultar en pérdida de rendimiento en términos de clasificación [16].

- Métodos de submuestreo: Consisten en descartar muestras de la clase mayoritaria de acuerdo a algún criterio. El método más simple es el submuestreo aleatorio (RUS), que implica la eliminación aleatoria de los ejemplos de la clase mayoritaria [17] hasta balancear la base de datos. Recientemente han sido implementados algoritmos más avanzados que hacen dicha eliminación basada en agrupamiento (e.g. k-means) [18] o basados en distancias (e.g. vecinos más cercanos) [19]. Los métodos de submuestro basados en agrupamiento buscan seleccionar las muestras más representativas de la clase

mayoritaria, particionando la base de datos en un número k de grupos usando algoritmos de *clustering*. Una vez se ha hecho esto, se selecciona un número adecuado de muestras mayoritarias de cada conglomerado considerando la relación entre el número de muestras minoritarias y mayoritarias en cada grupo. Por su parte, los métodos basados en distancia establecen algunas reglas de selección de muestras de la clase mayoritaria, de tal forma que se preserven aquellas observaciones cuya distancia promedio a una cantidad de muestras más cercanas de la clase minoritaria son las más pequeñas, o por el contrario, seleccionan los ejemplos cuya distancia promedio a las l -muestras de la clase minoritaria más lejanas son las más pequeñas.

El submuestreo en comparación con los métodos de sobremuestreo, usualmente ofrece un mayor rendimiento en tareas de clasificación, sin embargo por la necesidad de balancear la base de datos se pierde mucha información y se modifica la distribución original de los datos, lo que podría ocasionar sobre-entrenamiento [3].

- Métodos Híbridos: Estos son combinaciones entre métodos de sobremuestreo y submuestreo.

3.3.2. Aprendizaje costo-sensitivo

Mientras que los métodos de muestreo intentan equilibrar las distribuciones al considerar las proporciones representativas de ejemplos de clase en la distribución, los métodos de aprendizaje sensibles al costo consideran los costos asociados con clasificación errónea de las muestras [20]. En lugar de crear distribuciones equilibradas de datos a través de diferentes estrategias de muestreo, este tipo de aprendizaje se enfoca en el problema de aprendizaje desbalanceado usando diferentes matrices de costos que describen los costos de clasificar erróneamente cualquier ejemplo de datos particular. Por lo tanto, las técnicas costo-sensitivas proporcionan una alternativa viable a los métodos de muestreo para dominios de clasificación de datos desbalanceados.

Muchos de estos métodos están basados en modificación los umbrales de decisión o asignando pesos a nuevas instancias remuestreadas de acuerdo con la matriz de decisión de costos [21], [22]. Otra tipo de enfoque consisten en la manipulación de la función de costo de algunos algoritmos de aprendizaje de máquina como las máquinas de vectores de soporte (SVM) usando una estrategia de ponderación [23] o introduciendo una función de error sensible al costo en una red neuronal [24].

Estos clasificadores intentan aprender más características de la clase minoritaria estableciendo un alto costo a los errores de clasificación de dichas muestras. Sin embargo, dichos costos de error a menudo son desconocidos lo que puede resultar en sobre entrenamiento y en muchas ocasiones al dar prioridad a la clase minoritaria, ponen en peligro la especificidad lo que aumenta la cantidad de falsos positivos [3].

3.3.3. Métodos de ensamble

Los clasificadores basados en ensambles, son también denominados múltiples sistemas clasificadores [25] y mejoran el rendimiento de un solo clasificador combinando varios clasificadores base que superan a cada uno independiente [7]. Estos métodos están divididos en dos categorías; ensambles secuenciales y ensambles paralelos.

- Ensamblados secuenciales: Boosting es el más común y más efectivo método de ensamble [26]. El primer algoritmo de boosting aplicado al desbalance de clases fue Adaboost. En este tipo de ensambles las muestras que no se asignan a la clase correcta reciben pesos más altos, lo que obliga a un futuro clasificador a concentrarse más en el aprendizaje de estas muestras clasificadas fallidas.
- Ensamblados paralelos: Se refiere a modelos de ensamble en los cuales cada clasificador base puede ser entrenado en paralelo. Uno de los modelos más implementados es el denominado *Bagging*, el cual consiste en construir un conjunto impar de N clasificadores, cada uno entrenado en todas las instancias de clases minoritarias y N muestras de tamaño igual de instancias de clase mayoritaria, seleccionadas aleatoriamente. Cuando se clasifica una nueva instancia, cada clasificador entrenado hace una predicción, y la predicción final se toma como el voto mayoritario [8].

Si bien estos métodos suelen mostrar buenos resultados tienen la desventaja de tener alto costo computacional asociado a entrenar diversas máquinas de clasificación [3].

3.4. Medidas de rendimiento de clasificación para el desbalance de clases

Tradicionalmente, las medidas de rendimiento más frecuentemente usadas son la efectividad (*accuracy*) y la tasa de error (*error rate*). Considerando un problema básico de clasificación binaria, sean $\{p, n\}$ las clases positiva y negativa. Luego, una representación del rendimiento de clasificación puede ser formulado a través de una *matriz de confusión*, como se ilustra en la Tabla 3.1.

En este documento, se considera la clase minoritaria como la clase positiva y la clase mayoritaria como la negativa. Siguiendo esta convención, la precisión y la tasa de error son definidas como

		Valor real		total
		p	n	
Predicción	p'	Verdaderos Positivos (VP)	Falsos Positivos (FP)	N'_p
	n'	Falsos Negativos (FN)	Verdaderos Negativos (VN)	N'_n
total		N_p	N_n	

Tabla 3.1: Matriz de confusión

$$\text{Efectividad} = \frac{VP + VN}{VP + VN + FP + FN}; \quad \text{Tasa de error} = 1 - \text{Efectividad} \quad (3.1)$$

Estas medidas proveen una forma simple de describir el rendimiento de un clasificador sobre un determinado conjunto de datos. Sin embargo, pueden ser engañosas en ciertas situaciones y son muy sensibles a los cambios en los datos [3]. Suponga la situación de un conjunto de datos desbalanceado donde la clase mayoritaria representa el 95 % de las muestras y la minoritaria tan solo el 5 % de las observaciones. Un clasificador que prediga que todas las muestras pertenecen a la clase mayoritaria proveerá una efectividad del 95 %, lo cual parece un rendimiento excelente. Sin embargo, en la misma medida, esta descripción no refleja el hecho de que se identifica el 0 % de las muestras de la clase minoritaria. Es decir, la métrica de efectividad en este caso no proporciona información adecuada sobre la funcionalidad de un clasificador con respecto al tipo de clasificación requerida.

En lugar de la efectividad, otras métricas de desempeño se adoptan con frecuencia para proporcionar evaluaciones de problemas de aprendizaje con datos desbalanceados. Estas son, *Precision*, *Recall*, *F-measure* y *G-mean*. Estas métricas son definidas como:

$$\text{Precision} = \frac{VP}{VP + FP} \quad (3.2)$$

$$\text{Recall} = \frac{VP}{VP + FN} \quad (3.3)$$

$$\text{F-measure} = \frac{(1 + \beta)^2 \cdot \text{Recall} \cdot \text{Precision}}{\beta^2 \cdot \text{Recall} + \text{Precision}}, \quad (3.4)$$

donde β es un coeficiente para ajustar la importancia relativa de *Precision* contra *Recall* (usualmente $\beta = 1$).

$$\text{G-mean} = \sqrt{\frac{VP}{VP + FN} \times \frac{VN}{VN + FP}} \quad (3.5)$$

Aunque *F-measure* y *G-mean* son grandes mejoras sobre la efectividad, siguen siendo ineficaces de responder a preguntas más genéricas sobre las evaluaciones de clasificación [3].

Para superar tales dificultades, también se considera el análisis de la curva ROC (*Receiver Operating Characteristics* curves). Este análisis hace uso de las proporciones entre dos medidas de evaluación; la tasa de verdaderos positivos y la tasa de falsos positivos. Esta técnica genera curvas en lugar de una simple medida. Sin embargo, para comparar dos o más clasificadores se suele usar el área bajo la curva ROC (*AUC- Area under ROC Curve*) como criterio de evaluación.

Capítulo 4

Métodos de remuestreo

En un escenario de clasificación binaria de datos desbalanceados, una función de discriminación $g : \mathcal{X} \rightarrow \mathcal{Z}$ se aprende a partir de un conjunto $\{\mathbf{x}_n, z_n\}_{n=1}^N$, donde $\mathbf{x}_n \in \mathcal{X} \subseteq \mathbb{R}^P$ es un vector de entrada de características P -dimensional, correspondiente a la n -ésima muestra con etiqueta de salida $z_n \in \mathcal{Z} \subseteq \{-1, +1\}$. Se define también $\mathbf{X}_+ \in \mathbb{R}^{N_+ \times P}$ como los datos pertenecientes a la clase mayoritaria $z = +1$ y $\mathbf{X}_- \in \mathbb{R}^{N_- \times P}$ como las muestras pertenecientes a la clase minoritaria $z = -1$ con $N_+ \gg N_-$. El uso de métodos de remuestreo en aplicaciones de aprendizaje desbalanceado consiste en la modificación del conjunto de datos por algunos mecanismos con el fin de proporcionar una distribución equilibrada [8], lo que significa que $N_- \equiv N_+$ después de la operación.

En este capítulo se abordan algunos de los métodos de remuestreo más utilizados para el aprendizaje a partir de datos desbalanceados.

4.1. SMOTE

La técnica de creación de muestras sintéticas de la clase minoritaria conocida como SMOTE, es un método que ha mostrado tener éxito en diversas aplicaciones que involucran bases de datos desbalanceadas [27]. El algoritmo SMOTE crea datos artificiales entre muestras de la clase minoritaria. Específicamente, para el subconjunto $X_- \in X$, considere los k -vecinos más cercanos de cada muestra $\mathbf{x}_i \in X_-$, para algún entero k ; los k -vecinos más cercanos son definidos como las k muestras de X_- cuya distancia euclídea entre ellos y la muestra \mathbf{x}_i bajo consideración presenta las menores magnitudes. Para crear una muestra sintética, se selecciona aleatoriamente uno de los k -vecinos, luego se multiplica el correspondiente vector de diferencia por un número aleatorio entre el rango $[0, 1]$, y finalmente, se suma el anterior

resultado al vector \mathbf{x}_i

$$\mathbf{X}_{new} = \mathbf{x}_i + (\hat{\mathbf{x}}_i - \mathbf{x}_i) \times \delta, \quad (4.1)$$

donde $\hat{\mathbf{x}}_i$ es uno de los k -vecinos más cercanos de \mathbf{x}_i y $\delta \in [0, 1]$ es un número aleatorio. A pesar de que ha mostrado resultados positivos, el algoritmo SMOTE también tiene inconvenientes, incluyendo sobre-generalización y varianza [28].

4.2. Submuestreo basado en agrupamiento

Los algoritmos de muestreo basados en agrupamiento (CBUS - *Clustering-based Under-Sampling*) son particularmente interesantes porque proporcionan un elemento adicional de flexibilidad que no está disponible en la mayoría de los algoritmos de muestreo simples, y por lo tanto se puede adaptar a problemas muy específicos. Existen diversas variaciones para el método de submuestreo basado en agrupamiento. Sin embargo, en este documento abordamos dos de las más usadas.

4.2.1. Variante 1

En esta variante [6], primero se agrupan todas las muestras de entrenamiento en K *clusters*. La idea principal es que existen diferentes *clusters* en una base de datos, y cada grupo parece tener distintas características. Si un *cluster* tiene más muestras de clase mayoritaria y menos muestras de clase minoritaria, se comportará como las muestras de clase mayoritaria. Por otro lado, si un grupo tiene más muestras de clase minoritaria y menos muestras de clase mayoritaria, no tiene las características de las muestras de clase mayoritaria y se comporta más como las muestras de clase minoritaria. De acuerdo a esta información, esta variante selecciona un número adecuado de muestras de clase mayoritaria de cada grupo considerando la relación del número de muestras de la clase mayoritaria y el número de muestras de la clase minoritaria en cada *cluster*.

Inicialmente, se agrupa el conjunto de datos en K *clusters*. Una vez que los datos han sido agrupados, se selecciona al azar un número adecuado ($N_+^{*(k)}$) de muestras de clase mayoritaria del k -ésimo grupo, considerando la relación entre el número de muestras de clase mayoritaria y el número de muestras de clase minoritaria en cada grupo como sigue:

$$N_+^{*(k)} = (r \times N_-) \times \frac{N_+^{(k)} / N_-^{(k)}}{\sum_{k=1}^K N_+^{(k)} / N_-^{(k)}}, \quad (4.2)$$

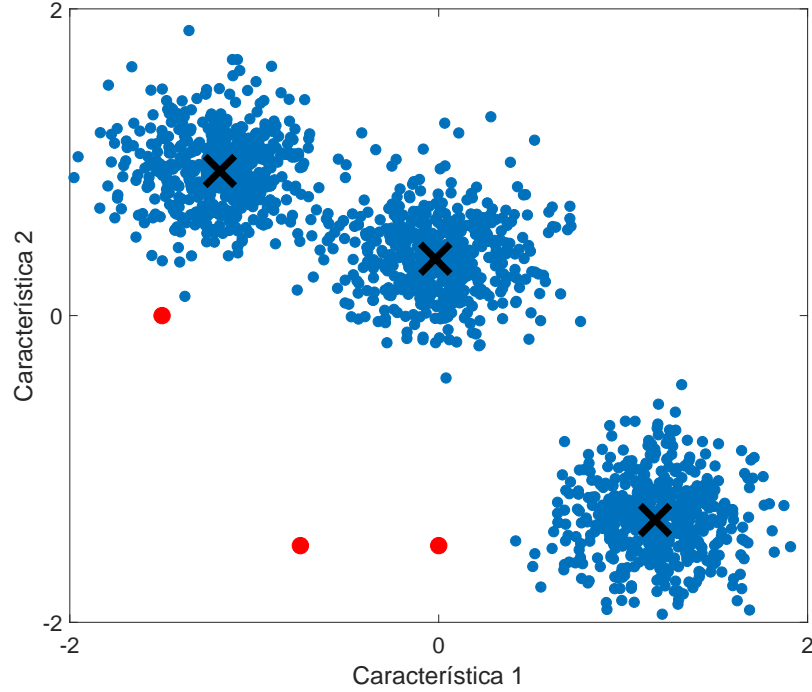


Figura 4.1: Ejemplo del CBUS: Las muestras azules representan la clase mayoritaria, las rojas representan a las de la clase minoritaria, y las equis representan las observaciones seleccionadas por el método de muestreo.

Donde r es la relación de desbalance esperada una vez se ha submuestreado la clase mayoritaria. Este método requiere de la sintonización empírica del número de *cluster* en los cuales se desea agrupar la base de datos, lo cual representa una desventaja para su correcta implementación.

4.2.2. Variante 2

Para evitar la sintonización del número de *clusters*, recientemente se publicó una nueva variante donde este asunto no es un problema [9]. En este método, la clase mayoritaria se agrupa en N_- , donde N_- es el número de muestras de clase minoritaria. Luego, los datos originales en los mismos grupos son reemplazados por los centros de cada *cluster*. Este procedimiento garantiza que el número de muestras de la clase mayoritaria sea reducido y balanceado con el número de muestras de la clase minoritaria.

En la Figura 4.1 se puede ver el funcionamiento del algoritmo CBUS en su segunda variante. En este caso, solo hay tres muestras de la clase minoritaria (clase roja), por lo tanto, siguiendo el procedimiento anteriormente descrito, primero se agrupa la clase mayoritaria en tres *clusters* (de acuerdo al número de muestras de la clase minoritaria), luego, los puntos en el mismo conglomerado son remplazados por el centroide de cada grupo, submuestreando la clase mayoritaria obteniendo así el mismo número de muestras en ambas clases.

4.3. Discusión y consideraciones de los métodos de remuestreo

- Los métodos de submuestreo permiten balancear un conjunto de datos desbalanceado a través de la remoción de muestras de la clase mayoritaria. Sin embargo, dichas técnicas descartan enormes cantidades de datos. Esto puede ser problemático, ya que la pérdida de tales datos puede hacer que las fronteras de decisión entre las instancias minoritarias y mayoritarias sea más difícil de aprender, lo que resulta en una pérdida en el rendimiento de clasificación [29]. Por lo tanto, en casos de niveles de desbalance muy altos se hace necesario de métodos que codifiquen la estructura general de la clase mayoritaria. En el caso contrario, los métodos de sobremuestreo crean muestras de maneras sintéticas de acuerdo a diferentes reglas, sin embargo, se corre con el riesgo de introducir ruido a los clasificadores debido a que las muestras nuevas no hacen parte del conjunto inicial de datos.
- Un inconveniente de las técnicas de remuestreo es que se necesita determinar cuánto muestreo se debe aplicar. En muchos casos, se debe elegir un nivel de muestreo excesivo para balancear ambas clases. De manera similar, se debe elegir un nivel de submuestreo para retener la mayor cantidad de información posible sobre la clase mayoritaria, al tiempo que se reducen los efectos del desbalance.
- El algoritmo SMOTE, presenta un problema conocido como generalización excesiva, y se atribuye en gran medida a la forma en como SMOTE crea muestras sintéticas. Específicamente, SMOTE genera la misma cantidad de muestras de datos sintéticos para cada muestra de la clase minoritaria y lo hace sin tener en cuenta las muestras vecinas de la clase mayoritaria, lo que aumenta la ocurrencia de la superposición entre clases [28].
- Los algoritmos de submuestreo basados en agrupamiento, poseen la dificultad de la inicialización de los centroides en los algoritmos de agrupamiento básicos como el k -means.

Capítulo 5

Métodos de ensamble

A diferencia de los métodos de aprendizaje automático ordinarios (que usualmente generan un solo clasificador), los métodos de ensamble entrenan a un conjunto de clasificadores básicos a partir de los datos de entrenamiento para hacer predicciones con cada uno de ellos, y luego combinan estas predicciones para tomar la decisión final. Algunos métodos de ensamble tienen la capacidad de impulsar (*Boost*) a los clasificadores con un rendimiento ligeramente mejor que simplemente escoger muestras al azar para formar cada clasificador, y con una capacidad de generalización sólida. Por lo tanto, los *aprendices base* a menudo son referidos como *aprendices débiles*. Esto también indica que en los métodos de ensamble, los *aprendices base* pueden tener una capacidad de generalización débil. En realidad, la mayoría de los algoritmos de aprendizaje, como los árboles de decisión, redes neuronales u otros métodos de aprendizaje automático, pueden ser usados como *aprendices de base*, y algún método de ensamble pueden aumentar su rendimiento. De acuerdo con la forma en que se generan los aprendices de base, los métodos de ensamble se pueden clasificar en dos subconjuntos: métodos de ensamble paralelos y métodos de ensamble secuencial. Los métodos de ensamble paralelo generan clasificadores que se entrenan con muestras independientes y después se combinan para tomar una decisión. Uno de los ensambles paralelos más popular es el método conocido como *Bagging* [30]. Los métodos de ensamble secuenciales generan clasificadores que trabajan sucesivamente, es decir, un primer clasificador tiene influencia en la generación de los posteriores. Un exponente de este tipo de metodología es el método denominado en la literatura como AdaBoost [31].

Capítulo 6

Metodología de clasificación de datos desbalanceados a partir de submuestreo basado en agrupamiento

Con el fin de dar cumplimiento al objetivo uno de la presente investigación, se presenta una metodología de clasificación de datos desbalanceados la cual consiste en la combinación de dos técnicas del estado del arte: submuestreo basado en agrupamiento variante 1 (ver sección 4.2.1) y la técnica de ensamble *Bagging* (ver sección 5.1). En esta nueva propuesta se establece el valor del nivel de desbalance deseado r de la ecuación 4.2 como un valor entero mayor que uno. Este procedimiento permite seleccionar más muestras de la clase mayoritaria, disminuyendo la pérdida de información. Sin embargo, esto hace que la base de datos siga desbalanceada al nivel $r : 1$. Una vez hecho el submuestreo se procede a usar las observaciones seleccionadas para crear un ensamble tipo *Bagging*. De acuerdo a lo anterior, si hay r veces más muestras de la clase mayoritaria que de la minoritaria, se procede a construir r clasificadores $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_r$, cada uno entrenado con todas las observaciones de la clase minoritaria y r subconjuntos de igual tamaño de muestras de la clase mayoritaria aleatoriamente seleccionadas. Finalmente, para clasificar una nueva instancia \mathbf{x}_{new} , cada clasificador entrenado hace su respectiva predicción, y la etiqueta final se decide por mayoría de votos.

Este procedimiento es especial para bases de datos con gran cantidad de muestras y con grandes niveles de desbalance, ya que disminuye en gran manera la pérdida de información, debido a que el procedimiento de agrupamiento codifica la forma global de la clase mayoritaria, por lo que se puede disminuir el tamaño de la base de datos preservando la estructura general de los datos. Además, el desbalance que queda se compensa con la técnica de *Bagging* la cual permite crear clasificadores con datos balanceados aumentando la capacidad para discriminar entre ambas clases.

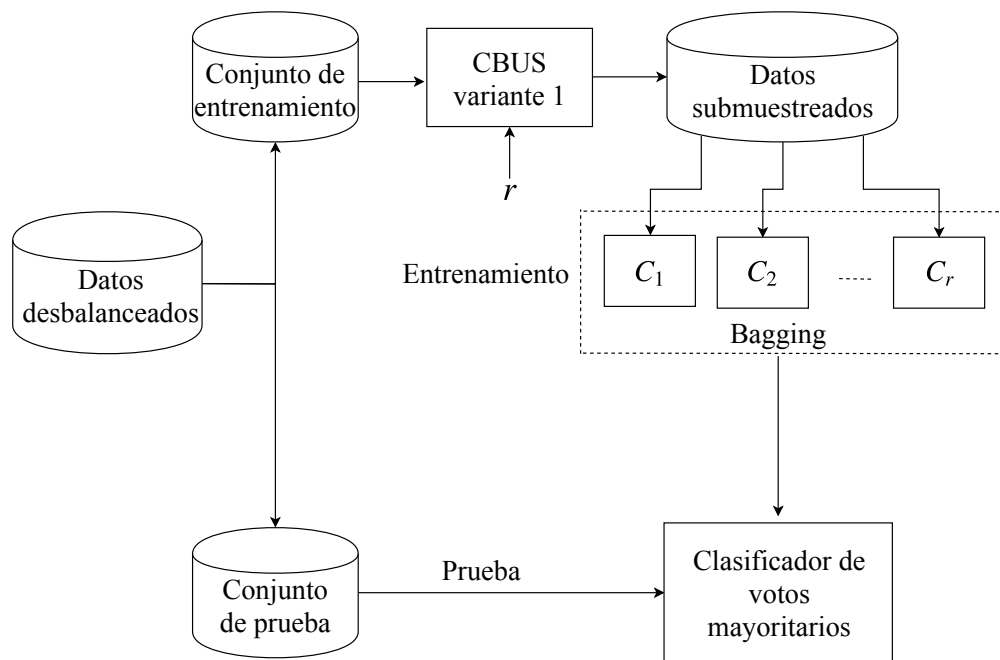


Figura 6.1: Diagrama metodología de clasificación de datos desbalanceados a partir del submuestreo basado en agrupamiento y la técnica *Bagging*.

Capítulo 7

Nueva propuesta de submuestreo basado en el principio de información relevante (RIS)

Como se ha mencionado anteriormente, el gran inconveniente de los métodos de submuestreo es la pérdida de información relevante debido a la eliminación aleatoria de muestras de manera total o parcial. Algunos métodos novedosos han intentado codificar la estructura de los datos para evitar remover muestras importantes, sin embargo, estos procedimientos están generalmente basados en técnicas sencillas de agrupamiento o en reglas creadas a partir de distancias euclídeas. Tales enfoques, si bien son una alternativa para los métodos de submuestreo aleatorio, y capturan la estructura global de los datos, no explotan al máximo la información de los mismos, ya que utilizan solamente estadísticos de hasta segundo orden (medias y varianzas).

Inspirados por el aprendizaje por teoría de la información (ITL-*Information Theoretic Learning*), se desea explotar las ventajas de los descriptores de entropía y disimilitud (divergencia e información mutua), que combinados con estimadores de funciones de densidad de probabilidad (PDF - *Probability density Function*) no paramétricos aportan solidez y generalidad a diversos métodos y mejoran el rendimiento en muchos escenarios realistas [32].

Sea $\{\mathbf{x}_n, z_n\}_{n=1}^N$, un conjunto desbalanceado de datos, donde $\mathbf{x}_n \in \mathcal{X} \subseteq \mathbb{R}^P$ es un vector de entrada de características P -dimensional, correspondiente a la n -ésima muestra con etiqueta de salida $z_n \in \mathcal{Z} \subseteq \{-1, +1\}$. Se define $\mathbf{X}_+ \in \mathbb{R}^{N_+ \times P}$ como los datos de la clase mayoritaria $z = +1$ y $\mathbf{X}_- \in \mathbb{R}^{N_- \times P}$ como la clase minoritaria $z = -1$ con $N_+ \gg N_-$.

La nueva propuesta de submuestreo, relacionada al objetivo dos de esta investigación, tiene como fin revelar las M muestras más relevantes o informativas de la clase mayoritaria tal

que $M < N_+$, lo cual podría evitar resultados de clasificación sesgados a la clase mayoritaria, capturando la estructura local de los datos. En consecuencia, este método está basado en un principio de auto-organización conocido como el principio de información relevante (PRI - *Principle of Relevant Information*).

7.1. Submuestreo basado en el principio de información relevante

El PRI es un método del marco del aprendizaje por teoría de la información (ITL-*Information Theoretic Learning*) el cual captura la estructura subyacente de los datos (modos, curvas principales, aproximaciones de cuantización de vectores entre otros). Aunque el método no se ha utilizado en el marco del tratamiento de datos desbalanceados, en esta investigación se ha adaptado para encontrar las M muestras de un conjunto de datos, para nuestro interés la clase mayoritaria, tal que $M < N_+$. El objetivo es encontrar el subconjunto de las M muestras de la clase mayoritaria cuya entropía es mínima, mientras preserva diferentes niveles de detalle respecto a las muestras originales de la clase con más muestras [32]. Específicamente, esta nueva adaptación, que hemos definido submuestreo basado en información relevante (RIS - *Relevant Information-based sampling*), busca el subconjunto $\tilde{\mathbf{X}}_+ \in \mathbb{R}^{M \times P}$ que minimiza la siguiente función de costo:

$$J(\tilde{\mathbf{X}}_+) = \min_{\tilde{\mathbf{X}}_+} \left[H_\alpha(\tilde{\mathbf{X}}_+) + \lambda D_\alpha(\tilde{\mathbf{X}}_+ || \mathbf{X}_+) \right], \quad (7.1)$$

donde $H_\alpha(X)$ es conocida como la entropía- α de Renyi, D_α es la divergencia de Renyi de orden α y $\lambda \in \mathbb{R}^+$ es un parámetro de compensación que controla el nivel de distorsión de los datos de submuestreo. Este tipo de medidas hacen parte del marco del ITL, por lo que se recomienda ver Apéndice A.

Para propósitos de estimación, se propone medir la redundancia por medio de la entropía cuadrática de Renyi H_2 , y medir la distorsión entre la clase mayoritaria y las M muestras a procesar por medio de la Divergencia de Cauchy - Schwarz D_{CS} debido a la equivalencia mostrada en la ecuación (A.5) del Apéndice A. Esto conlleva a la siguiente expresión:

$$\begin{aligned} J(\tilde{\mathbf{X}}_+) &= \min_{\tilde{\mathbf{X}}_+} \left[H_2(\tilde{\mathbf{X}}_+) + \lambda D_{CS}(\tilde{\mathbf{X}}_+ || \mathbf{X}_+) \right] \\ &= \min_{\tilde{\mathbf{X}}_+} \left[(1 - \lambda) H_2(\tilde{\mathbf{X}}_+) + 2\lambda H_2(\tilde{\mathbf{X}}_+, \mathbf{X}_+) - \lambda H_2(\mathbf{X}_+) \right]. \end{aligned} \quad (7.2)$$

El último término en la ecuación (7.2) es constante con respecto a $\tilde{\mathbf{X}}_+$, por lo tanto, para la optimización se puede reducir la función de costo $J(\tilde{\mathbf{X}}_+)$ a

$$\begin{aligned} J(\tilde{\mathbf{X}}_+) &= \min_{\tilde{\mathbf{X}}_+} \left[(1 - \lambda) \hat{H}_2(\tilde{\mathbf{X}}_+) + 2\lambda H_2(\tilde{\mathbf{X}}_+, \mathbf{X}_+) \right] \\ &= \min_{\tilde{\mathbf{X}}_+} \left[-(1 - \lambda) \log V(\tilde{\mathbf{X}}_+) - 2\lambda \log V(\tilde{\mathbf{X}}_+, \mathbf{X}_+) \right]. \end{aligned} \quad (7.3)$$

Diferenciando $J(\tilde{\mathbf{X}}_+)$ con respecto a $\{\tilde{\mathbf{x}}_{+k}\}_{k=1}^M$ e igualando a cero se obtiene:

$$\begin{aligned} \frac{\partial J(\tilde{\mathbf{X}}_+)}{\partial \tilde{\mathbf{x}}_{+k}} &= \frac{2(1 - \lambda)}{V(\tilde{\mathbf{X}}_+)} \frac{\partial V(\tilde{\mathbf{X}}_+)}{\partial \tilde{\mathbf{x}}_{+k}} + \frac{2\lambda}{V(\tilde{\mathbf{X}}_+, \mathbf{X}_+)} \frac{\partial V(\tilde{\mathbf{X}}_+, \mathbf{X}_+)}{\partial \tilde{\mathbf{x}}_{+k}} = 0 \\ &= \frac{2(1 - \lambda)}{V(\tilde{\mathbf{X}}_+)} F(\tilde{\mathbf{x}}_{+k}, \tilde{\mathbf{X}}_+) + \frac{2\lambda}{V(\tilde{\mathbf{X}}_+, \mathbf{X}_+)} F(\tilde{\mathbf{x}}_{+k}, \mathbf{X}_+) = 0. \end{aligned} \quad (7.4)$$

De acuerdo a esta última ecuación, las muestras de procesamiento $\tilde{\mathbf{X}}_+$ se moverán bajo la influencia de dos fuerzas: una fuerza de información $F(\tilde{\mathbf{x}}_{+k}, \mathbf{X}_+)$ ejercida por las demás muestras de $\tilde{\mathbf{X}}_+$ en proceso, y una fuerza de información cruzada $F(\tilde{\mathbf{x}}_{+k}, \mathbf{X}_+)$ ejercida por las muestras pertenecientes a la clase mayoritaria \mathbf{X}_+ . Finalmente, reordenando la ecuación (7.4) se obtiene una ecuación de actualización basada en el método de punto-fijo como se observa en la siguiente ecuación:

$$\begin{aligned} \tilde{\mathbf{x}}_{+k}^{(n+1)} &= c \frac{(1 - \lambda)}{\lambda} \frac{\sum_{j=1}^M G_\sigma(\tilde{\mathbf{x}}_{+k}^{(n)}, \tilde{\mathbf{x}}_{+j}^{(n)}) \tilde{\mathbf{x}}_{+j}^{(n)}}{\sum_{j=1}^{N_+} G_\sigma(\tilde{\mathbf{x}}_{+k}^{(n)}, \mathbf{x}_{+j})} + \frac{\sum_{j=1}^{N_+} G_\sigma(\tilde{\mathbf{x}}_{+k}^{(n)}, \mathbf{x}_{+j}) \mathbf{x}_{+j}}{\sum_{j=1}^{N_+} G_\sigma(\tilde{\mathbf{x}}_{+k}^{(n)}, \mathbf{x}_{+j})} \\ &\quad - c \frac{(1 - \lambda)}{\lambda} \frac{\sum_{j=1}^M G_\sigma(\tilde{\mathbf{x}}_{+k}^{(n)}, \tilde{\mathbf{x}}_{+j}^{(n)})}{\sum_{j=1}^{N_+} G_\sigma(\tilde{\mathbf{x}}_{+k}^{(n)}, \mathbf{x}_{+j})} \tilde{\mathbf{x}}_{+k}^{(n)}, \end{aligned} \quad (7.5)$$

donde $c = \frac{\hat{V}(\tilde{\mathbf{X}}_+, \mathbf{X}_+)}{\hat{V}(\tilde{\mathbf{X}}_+)} \frac{N_+}{M}$.

Un ejemplo ilustrativo es llevado a cabo para mostrar el funcionamiento del método RIS. En la figura 7.1 se observa una muestra bi-dimensional \mathbf{X}_+ de puntos distribuidos en una arreglo no-lineal. Suponga que la cara corresponde a la clase mayoritaria la cual necesitamos submuestrear. Lo que se desea es preservar los detalles de la cara tanto como sea posible. De acuerdo a esto, se emplea el algoritmo 2, por lo que primero se generan M muestras de $\tilde{\mathbf{X}}_+$ las cuales se empiezan a esparcir mientras minimizan la función de costo. Finalmente, se puede observar la solución alcanzada por el RIS (figura 7.1-d) el cual revela las M instancias más relevantes de la silueta en análisis, las cuales preservan la estructura general de la cara.

Algoritmo 2 Algoritmo de submuestreo basado en Información Relevante (RIS)

Entrada: \mathbf{X}_+ , σ and λ

- 1: Generar M muestras pertenecientes a $\tilde{\mathbf{X}}_+^{(0)}$
 - 2: Se calcula el potencial de información de $\tilde{\mathbf{X}}_+$ de acuerdo a la ecuación (A.4)
 - 3: Calcular el potencial de información cruzado entre $\tilde{\mathbf{X}}_+$ y \mathbf{X}_+ de acuerdo a la ecuación. (A.6).
 - 4: Actualizar las muestras de $\tilde{\mathbf{X}}_+$ con la regla de punto fijo de la ecuación. (7.5).
 - 5: Iterar los pasos 2 al 4 hasta la convergencia
-

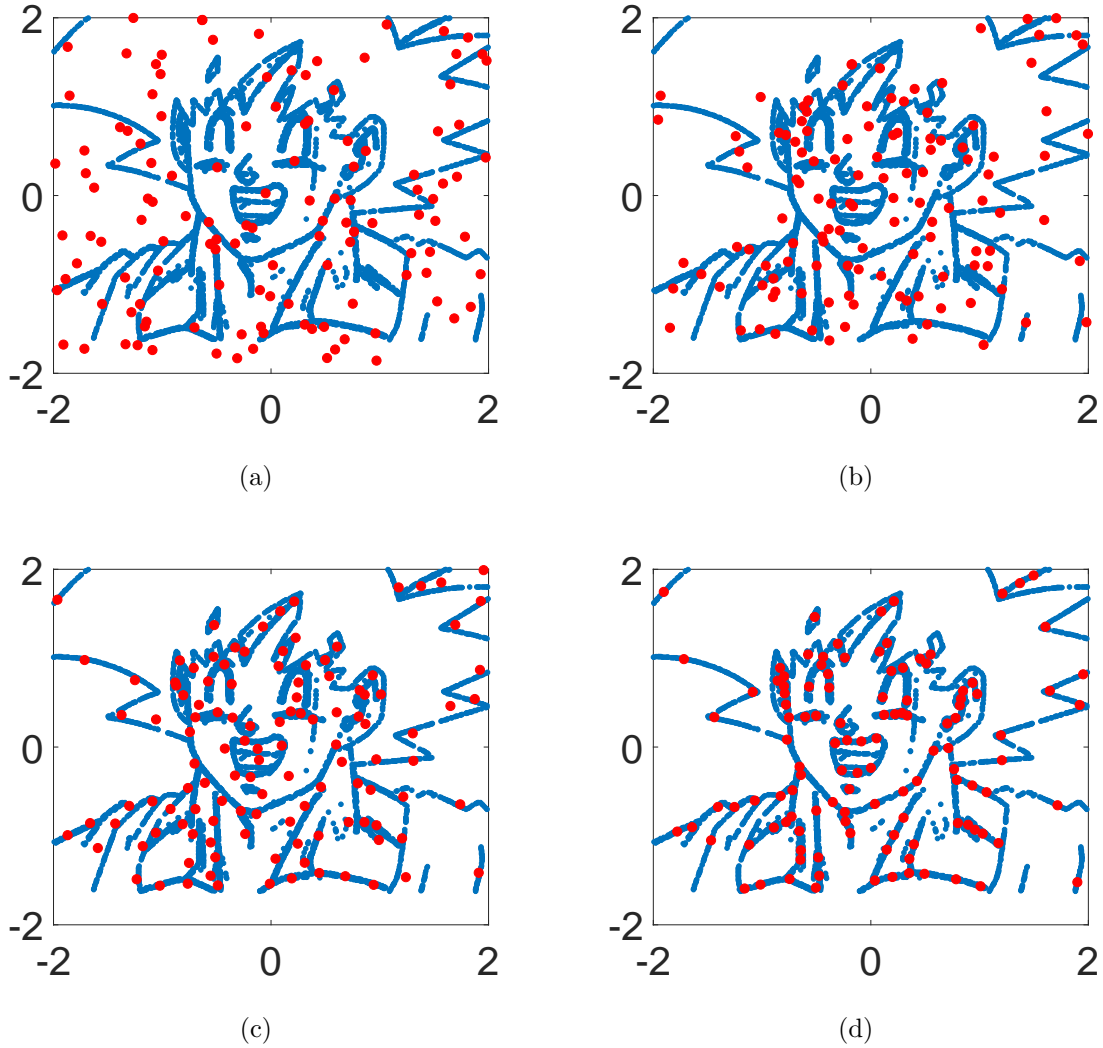


Figura 7.1: Ejemplo ilustrativo. Se muestra el funcionamiento del algoritmo en la (a) 1^{ra}, (b) 10^{ava}, (c) 50^{ava} y (d) 655^{ava} iteración.

7.2. Parámetros libres

Respecto a los parámetros libres del algoritmo RIS, su función de costo es descrito por solo dos parámetros: el parámetro de ponderación λ y el parámetro del ancho de banda σ del kernel en la estimación de Parzen. El parámetro de resolución controla la escala del análisis, mientras que el parámetro de ponderación combina los términos de regularización y similitud en proporciones adecuadas para capturar diferentes aspectos de la estructura de los datos. Como es usual en los métodos kernel, el principal asunto es como escoger un valor apropiado para σ que permita obtener buenos resultados. Por lo tanto, se implementó la estrategia propuesta por Principe et.al. [32], el cual reduce los riesgos de que la función de costo quede atrapada en un mínimo local y hace al algoritmo más flexible. La estrategia de aprendizaje consiste en disminuir el valor del σ a través de las iteraciones así: $\sigma_n = \frac{\sigma}{1+\gamma n}$, donde γ es una tasa de decaimiento. Esto significa que es necesario determinar el valor apropiado para el ancho de banda inicial σ , y la tasa de decaimiento.

Con el fin de encontrar los valores adecuados para σ , γ y λ en la sección 9.1.3 del marco experimental, se presentan algunas pruebas que permiten establecer dichos valores para obtener los mejores resultados.

Capítulo 8

CRIS: Combinación del RIS, la técnica de CBUS y el método de ensamble Bagging

Una importante implicación del uso de la técnica RIS es que para grandes cantidades de datos presenta problemas para el almacenamiento y cálculo de la matriz *kernel*, ya que su almacenamiento es del orden $O(N^2)$ y su cálculo requiere $O(N^2P)$ operaciones. En consecuencia, el cálculo del potencial de información y el potencial de información cruzado son del orden $O(M^2P)$ y $O(NMP)$, lo cual ocasiona costos computacionales muy altos para bases de datos con muchas muestras. Para superar esta limitación, se propone una mejora al método a través de la combinación de RIS y la técnica de submuestreo basado en agrupamiento (CBUS-variante 2).

El objetivo de CBUS es agrupar objetos similares (muestras de datos) en los mismos conglomerados; los objetos en diferentes grupos son diferentes en términos de sus representaciones en el espacio de características. Por lo tanto, CBUS genera una serie de *cluster*, y cada uno de estos contiene datos similares. Específicamente, en el método original, la clase mayoritaria se agrupa en N_- conjuntos, donde N_- es el número de muestras de la clase minoritaria. Luego, los datos originales de los mismos grupos son reemplazados por los centroides de cada *cluster*, lo que reduce el tamaño de la clase mayoritaria. Por lo tanto, se ha propuesto una variante de la técnica de CBUS al seleccionar (en lugar del centroide) las M muestras más relevantes de cada grupo usando el RIS minimizando la función de costo de la ecuación. (7.3).

Hay dos principales motivaciones para el uso de la técnica CBUS en combinación con el RIS. Primero, al agrupar la clase mayoritaria en N_- *clusters* implica que cada conglomerado tendrá en promedio tantas muestras como sea el nivel de desbalance, por ejemplo, un conjunto de datos con 1000 muestras de la clase mayoritaria y 100 muestras de la clase minoritaria tiene

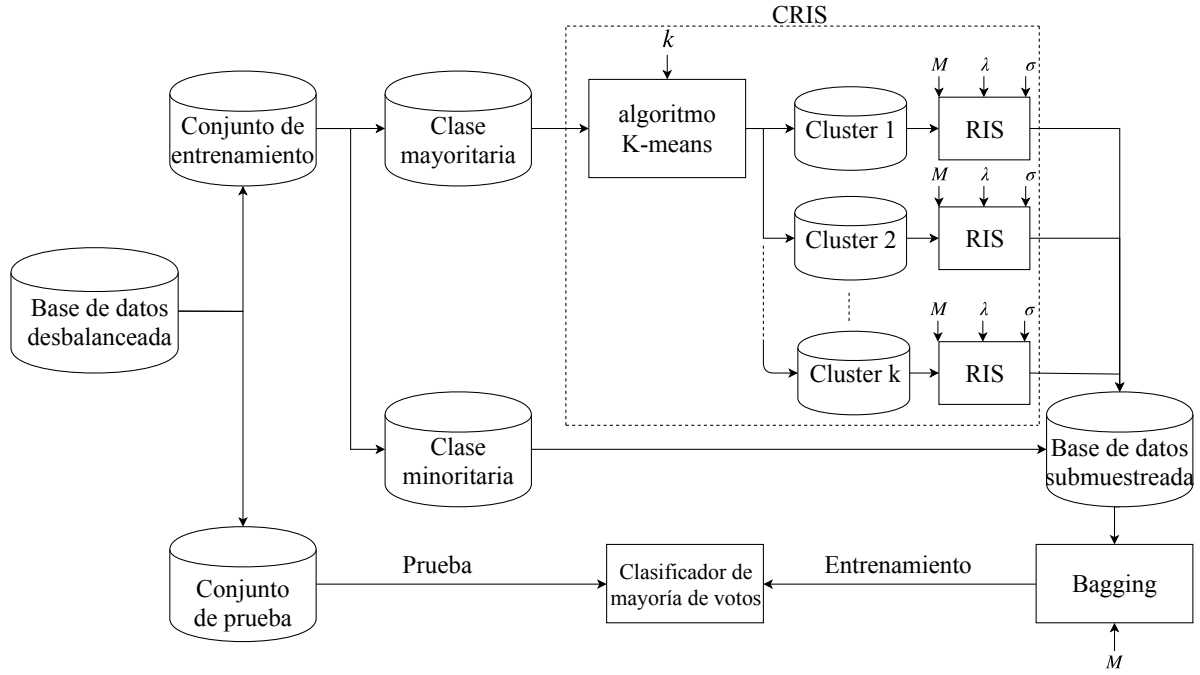


Figura 8.1: Diagrama metodología de submuestreo y clasificación de datos desbalanceados a partir de la técnica CRIS.

un orden de desbalance de 10:1, por lo tanto, si agrupamos la clase mayoritaria en $N_- = 10$ grupos cada grupo tendrá en promedio 10 muestras. Por lo anterior, la restricción del número de muestras se torna irrelevante. En segundo lugar, el algoritmo CBUS original selecciona el centroide de cada grupo, pero la media de un conjunto de datos puede ser un resultado sesgado si el *cluster* contiene datos atípicos (*outliers*) o si la forma del conglomerado no es Gaussiana. Por tal razón, el punto que selecciona CBUS podría no representar el grupo. En este sentido, se propone seleccionar, en lugar del centroide, M muestras de cada *cluster*, balanceando el conjunto de datos al orden $M : 1$, permitiendo construir un ensamble tipo *Bagging* como se explicó en la sección 5.1. Esta variación se ha denominado CRIS y en la Figura 8.1 se ilustra el funcionamiento de este nuevo enfoque de ensamble.

Parte III

Marco Experimental

Capítulo 9

Esquema de trabajo

En este capítulo se describe la configuración de las pruebas realizadas para la comparación las técnicas propuestas para tratar el desbalance de clases en diferentes tipos de datos. En los experimentos se contempla la clasificación de bases de datos de pequeña escala ($N < 5000$ muestras). Además, se explica el tipo y conformación de las bases de datos sobre las cuales se llevan a cabo las pruebas. Por otra parte, se evalúan las metodologías propuestas sobre bases de datos de gran escala para analizar la escalabilidad de los métodos en presencia de conjuntos de datos con enormes cantidades de datos.

Todos los algoritmos aplicados para las pruebas a continuación descritas fueran elaborados sobre la herramienta de programación Matlab ® y la implementación de los métodos propuestos se pueden encontrar libremente en un repositorio en línea ¹

9.1. Clasificación de bases de datos desbalanceadas

Con el fin de comparar el desempeño de RIS y CRIS como técnicas para tratar bases de datos desbalanceadas, se desarrollan algunos experimentos sobre bases de datos sintéticas y del mundo real.

¹<https://github.com/keider95/Relevant-Information-Sampling/>

9.1.1. Descripción de las bases de datos

Datos sintéticos

Primero, se lleva a cabo un experimento representativo para comparar el funcionamiento de RIS y CRIS contra dos técnicas comunes de submuestreo. Los datos sintéticos consisten en una muestra de puntos bidimensionales distribuidos en un arreglo de tres conglomerados con formas no Gaussianas, los cuales representan la clase mayoritaria y tres puntos correspondientes a la clase minoritaria.

Datos del mundo real

Bases de datos de pequeña escala Para los experimentos sobre bases de datos reales de pequeña escala, se usaron 44 bases de datos para clasificación binaria del repositorio de datos KEEL [8], las cuales están disponibles públicamente ². Estas bases de datos tienen órdenes de desbalance (OD) entre 1.8 y 129 y el número de muestras varía entre 130 y 5500. Una breve descripción de estos conjuntos de datos se muestra en la Tabla 9.1.

Bases de datos de gran escala Para evaluar la escalabilidad de los métodos ante bases de datos de gran escala, se prueban tres bases de datos de diferentes aplicaciones: dos del *Knowledge Discovery and Data Mining Cup* ³, estos son los conjuntos de datos de predicción de cáncer de mama y de homología de proteínas, que contienen 102294 y 145751 muestras de datos, 117 y 74 características y relaciones de desequilibrio de 163 y 111, respectivamente.

Una base de datos de una escala aún mayor es probada también. Se trata de una base de datos de detección automática de displasias corticales focales (FCD-*Focal Cortical Dysplasias*) usada en [33]. ⁴ La base de datos consiste en la caracterización de la corteza cerebral de 22 imágenes de resonancia magnética (MRI-*Magnetic Resonance Image*) cuyo número de características es $P = 28$ y el número total de muestras es 3307529. Finalmente, el orden del desbalance es del orden 42:1.

²<http://www.keel.es/dataset.php>

³<http://www.kdd.org/kdd-cup>.

⁴Disponible *online* en <https://doi.org/10.17863/CAM.6923>

Tabla 9.1: Información de las bases de datos de pequeña escala

	Nombre	N	P	OD		Nombre	N	P	OD
1	Abalone9-18	731	8	16.68	23	Page-blocks13vs2	472	10	15.85
2	Abalone19	4174	8	128.87	24	Pima	768	8	1.9
3	Ecoli-0_vs_1	220	7	1.86	25	Segment0	2308	19	6.01
4	Ecoli-0-1-3-7_vs_2-6	281	7	39.15	26	Shuttle0vs4	1829	9	13.87
5	Ecoli1	336	7	3.36	27	Shuttle2vs4	129	9	20.5
6	Ecoli2	336	7	5.46	28	Vehicle0	846	18	3.23
7	Ecoli3	336	7	8.19	29	Vehicle1	846	18	2.52
8	Ecoli4	336	7	13.84	30	Vehicle2	846	18	2.52
9	Glass0	214	9	3.19	31	Vehicle3	846	18	2.52
10	Glass0123vs456	192	9	10.29	32	Vowel0	988	13	10.1
11	Glass016vs2	184	9	19.44	33	Wisconsin	683	9	1.86
12	Glass016vs5	214	9	1.82	34	Yeast05679vs4	528	8	9.35
13	Glass1	214	9	10.39	35	Yeast1	1484	8	2.46
14	Glass2	214	9	15.47	36	Yeast1vs7	459	8	13.87
15	Glass4	214	9	22.81	37	Yeast1289vs7	947	8	30.56
16	Glass5	214	9	22.81	38	Yeast1458vs7	693	8	22.1
17	Glass6	214	9	6.38	39	Yeast2vs4	514	8	9.08
18	Haberman	306	3	2.68	40	Yeast2vs8	482	8	23.1
19	Iris0	150	4	2	41	Yeast3	1484	8	8.11
20	New-thyroid1	215	5	5.14	42	Yeast4	1484	8	28.41
21	New-thyroid2	215	5	4.92	43	Yeast5	1484	8	32.78
22	Page-blocks0	5472	10	8.77	44	Yeast6	1484	8	39.15

9.1.2. Descripción de las pruebas de submuestreo

Submuestreo basado en información relevante (RIS)

Antes de submuestrear las bases de datos, se emplea una normalización *Z-score* para todos los datos de entrenamiento, convirtiendo todas las características a una escala estándar con media cero y desviación uno. Luego, para correr el algoritmo RIS, se toma la clase mayoritaria \mathbf{X}_+ de cada conjunto de datos, y se colocan M muestras pertenecientes a $\tilde{\mathbf{X}}_+$ en posiciones aleatoria dentro de un hipercubo P -dimensional en el rango -2 a 2, donde, después de haber normalizado, se puede encontrar aproximadamente el 95 % de los datos. El Número M de puntos para seleccionar de la clase mayoritaria es establecido como igual al número de muestras de la clase minoritaria (i.e. $M = N_-$); en consecuencia, tanto al clase mayoritaria como la clase minoritaria contendrán la misma cantidad de muestras. Finalmente, usamos la ecuación (7.5), permitiendo a las partículas auto-organizarse, minimizando la función de costo expresada por la ecuación 7.2 hasta la convergencia. El algoritmo converge cuando la diferencia entre la salida del algoritmo de una iteración a otra presenta un cambio mínimo, esto es $\|\tilde{\mathbf{X}}_+^{(n)} - \tilde{\mathbf{X}}_+^{(n-1)}\| < \epsilon$.

Submuestreo basado en agrupamiento e información relevante (CRIS)

Para emplear este método de submuestreo, se sigue el procedimiento descrito en la figura 8.1. Primero, se separa la clase mayoritaria de la minoritaria. Luego, se agrupa la clase mayoritaria en N_- clusters, donde N_- es el número de muestras de la clase minoritaria. Después de esto, se aplica el método de RIS sobre cada conglomerado, seleccionando las M muestras más relevantes de cada cluster, donde $M \in \mathbb{R}^+ \geq 1$. Hecho este proceso, la base de datos quedará submuestreada al orden $M : 1$, es decir, la clase mayoritaria tendrá $M \times N_-$ muestras. Para los experimentos, se propone usar un valor de $M > 1$ con la intención de formar un ensamble como se explica en la sección 8, ya el número M de muestras determina el número de clasificadores a ensamblar.

9.1.3. Sintonización de parámetros libres

Como se menciona en la sección 7.2 respecto a los parámetros libres del RIS, la función de costo de la ecuación 7.3 es descrita por solo dos parámetros libres: el parámetro de ponderación λ y el parámetro de resolución σ . Como es usual en métodos *kernel*, el asunto es cómo escoger el valor de σ que permita obtener buenos resultados; en consecuencia, se implementa la estrategia de aprendizaje propuesta por Principe et.al. [32]. La estrategia consiste en permitir al ancho de banda disminuir a medida que pasan las iteraciones de acuerdo a $\sigma_n = \frac{\sigma}{1+\gamma n}$,

donde γ es una tasa de decaimiento. Esto significa que un valor inicial del de σ y la tasa de decaimiento necesitan ser determinadas para el problema. Esta metodología tiene como fin evitar que el método quede atrapado en un mínimo local y hace al algoritmo robusto y flexible.

Para encontrar los valores apropiados para estos parámetros, se desarrollan algunos experimentos en una base de datos sintética la cual consiste en dos grupos de puntos provenientes de dos distribuciones normales bivariadas con diferentes medias, modaradamente solapadas. En consecuencia, se establece fijo $\gamma = 1$ y se evalúan 7 diferentes tamaños iniciales del kernel: $\sigma \in \{\sigma_0, 2\sigma_0, 5\sigma_0, 10\sigma_0, 20\sigma_0, 50\sigma_0, 100\sigma_0\}$ donde σ_0 es la mediana de la distancia de los datos. La Figura 9.1(a) muestra la función de costo y el número de iteraciones en las cuales el método converge para los 7 valores de escala del kernel. De acuerdo a esto, se puede ver que entre más pequeño σ , más rápido el algoritmo converge. Este hallazgo demuestra que es apropiado establecer $\sigma = \sigma_0$ ya que asegura una rápida y estable convergencia. Sin embargo, es importantes resaltar que un valor de σ más pequeño que σ_0 podría hacer al algoritmo inestable y evitaría que el algoritmo converja. Por lo anterior, se recomienda fijar el parámetro de resolución del kernel inicial como σ_0 .

Respecto a λ , se compararon las soluciones alcanzadas por diferentes valores de ese parámetro: $\{\lambda \in \mathbb{N} | 0 < \lambda \leq 100\}$. En la figura 9.1(b) se puede observar la comparación de diferentes submuestreos de la clase mayoritaria, dado por cada valor de λ , en términos del AUC (AUC-*Area Under ROC Curve*, ver 3.4). De acuerdo a esta figura, se puede ver que para pequeños valores de λ el desempeño de clasificación es muy variable, pero después de $\lambda = 30$ el valor de AUC se estabiliza ya que para $\lambda \geq 30$ las soluciones alcanzadas fue casi la misma. De acuerdo a lo anterior, se decide escoger un valor grande de λ , concretamente $\lambda = 100$, ya que a medida que el valor de λ incrementa, el énfasis en la función de costo es puesto sobre la medida de similitud D_{CS} obteniendo submuestreos tan similares como sea posible a la clase mayoritaria.

9.1.4. Descripción de las pruebas de clasificación

Pruebas de clasificación sobre las bases de datos de pequeña escala

Para examinar el rendimiento de clasificación de las metodologías propuestas, se implementa cuatro diferentes clasificadores; Análisis Discriminante Lineal (LDA - *Linear Discriminant analysis*), K-vecinos más cercanos (KNN - *K-nearest neighbor*), máquinas de vectores de soporte (SVM - *Support Vector Machines*) y perceptron multicapa (MLP - *Multilayer Perceptron*). Los parámetros libres de cada uno de los clasificadores (el número K de vecinos, el parámetro de regularización y el tamaño del kernel en la SVM, y el número de capas y neuronas del MLP) fueron ajustados usando una estrategia de cros-validación anidada.

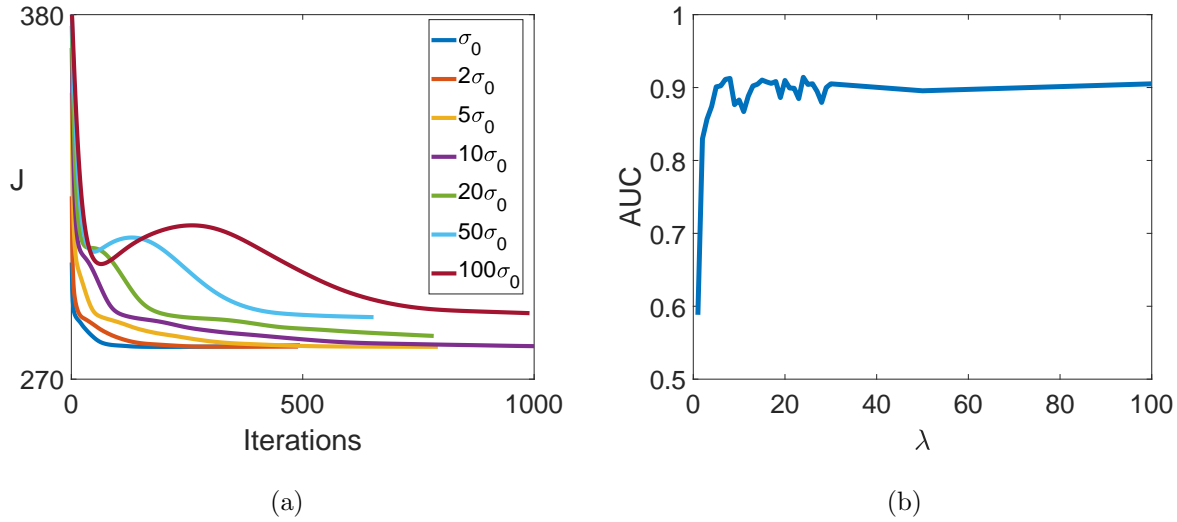


Figura 9.1: (a) Número de iteraciones hasta la convergencia, y (b) AUC alcanzadas por diferentes valores de λ .

También, se evalúa el desempeño en la clasificación como el AUC usando un esquema de cros-validación de 5 *fold*. Esto es, la base de datos se divide aleatoriamente en 5 grupos. Para cada grupo, los algoritmos son entrenados con las muestras de los 4 grupos restantes y luego se prueba el clasificador sobre el conjunto actual. las particiones usadas en este estudio pueden ser encontradas en los repositorios de las bases de datos, de tal manera que cualquier investigador interesado pueda reproducir los experimentos.

Para las bases de datos de predicción de cáncer de mama y de homología de proteínas se escoge el clasificador con mayor rendimiento en el análisis anteriormente descrito.

9.1.5. Pruebas sobre la base de datos de FCD

La base de datos de detección automática de displasias requiere un tratamiento diferente, ya que esta posee una mayor cantidad de muestras que las bases de datos antes mencionadas. Para tratar con esta base de datos se proponen dos técnicas: submuestro basado en *clustering* variante 1 (CBUS-1), y la combinación de CBUS-1 y la técnica *bagging* ya mencionada. Para implementar CBUS-1 es necesario fijar un número de *cluster* para agrupar la base de datos, el cual se establece de manera heurística. Para este caso, los datos son agrupados en $K = 4$ conglomerados. Para implementar esta primera técnica se fija el orden de desbalance esperado como uno ($r = 1$ de la ecuación 4.2) balanceando al rodén 1:1. Luego, CBUS-1 es implementado por cada paciente, y los datos submuestreados son concatenados resultando en una matriz de entrada $\tilde{\mathbf{X}}$ con $n = 151340$ muestras, $p = 28$ características y con el mismo

número de muestras en ambas clases. Para combinar CBUS-1 con *bagging* se establece el orden de desbalance esperado como $r = 5$, balanceando al orden 5:1. Esto permite entrenar 5 clasificadores paralelamente, cada uno entrenado con todas las muestras correspondientes a displasias y 5 muestras de igual tamaño con las muestras más relevantes de la clase sana.

Para este caso, se comparan los resultados logrados contra las metodologías implementadas en el estado del arte para esta base de datos [33–35]: sin submuestreo (WUS - *without undersampling*), submuestreo aleatorio (RUS - *Random Undersampling*), y *bagging* con RUS. Para todos los esquemas se entrenaron solo redes neuronales. Se elige una red neuronal de una sola capa oculta como clasificador porque se puede entrenar rápidamente en conjuntos con gran cantidad de datos [33]. El número de neuronas ocultas en la red se determina a través de un análisis de componentes principales aplicado a los datos de entrada. Se establece el número de neuronas como el número de componentes que explican 99% de la varianza. En este caso, se requieren 11 componentes principales. Finalmente, empleamos un análisis de validación cruzada de 10 veces para probar el rendimiento de los métodos. El *G-mean*, la sensibilidad y la especificidad de los resultados de los clasificadores se calculan porque estas medidas nos proporcionan una estimación de rendimiento clase por clase. Las curvas ROC también se emplean para evaluar el rendimiento de cada enfoque. Finalmente, se comparan los resultados globales de cada método.

9.1.6. Métodos de comparación

Para las bases de datos de pequeña escala, se compara el rendimiento de RIS y CRIS contra cuatro métodos del estado del arte que ha mostrado superar los demás métodos para tratar con bases de datos desbalanceados [8]: *i) RUSBoost1* (RUS1): el cual es una combinación de RUS y el procedimiento estándar de *Adaboost* [36]. *ii) Underbagging4* (UB4): este es una combinación de RUS y el procedimiento de *bagging*. *iii) SMOTEBagging4* (SBAG4): Este método involucra un paso de generación de muestras sintéticas de la clase minoritaria (SMOTE) combinado con la construcción de un ensamble de tipo *bagging* [37]. *iv) (CBUS-AdaBoost)*: CBUS variante-2 combinado con *AdaBoost* [9].

Capítulo 10

Resultados de las pruebas sobras bases de datos de pequeña escala

A continuación, se presentan los resultados de todas las pruebas realizadas, sobre los datos sintéticos y cada una de las 44 bases de datos de pequeña escala.

10.1. Resultados base de datos sintética

Como se mencionó en el capítulo anterior, el primer experimento consiste en comparar el funcionamiento de RIS y CRIS contra dos técnicas comunes de submuestreo, a saber, el submuestreo aleatorio (RUS) y el submuestreo basado en agrupamiento variante 2. La figura 10.1 muestra una muestra bidimensional de puntos distribuidos en un arreglo no-lineal de 3 conglomerados que representan la clase mayoritaria. También, en la figura se pueden ver 3 puntos rojos que corresponden a la clase minoritaria.

En esta base de datos lo que se desea es seleccionar tres puntos de la clase mayoritaria para balancear el conjunto de datos, pero preservando la estructura global de los datos. La figura 10.1 (a) corresponde a las muestras seleccionadas por RUS. Como se puede ver, RUS selecciona una muestra del *cluster* izquierdo, y dos muestras del *cluster* central, sin embargo, el método selecciona las muestras al azar, por lo que las pudiera haber tomado de cualquier lugar. En este sentido, debido a que RUS no tiene en cuenta la estructura global de los datos para submuestrear la clase mayoritaria, los puntos seleccionado no representan adecuadamente la clase con más datos. Por otra parte, se puede observar en la figura 10.1 (b) que CBUS agrupa la clase mayoritaria en tres *clusters* y el método selecciona el centroide (o la muestra más cercana al centroide) de cada grupo. También, note que tanto en el *cluster*

izquierdo como en el del medio, el centroide no es parte de las muestras originales. Por lo tanto, es claro que para *clusters* con formas no Gaussianas, la media no es un valor adecuado para representar el grupo. La figura 10.1 (c) y (d) muestra los puntos seleccionados por las técnicas RIS y CRIS. RIS busca las muestras más relevantes o informativas de la clase mayoritaria completa, mientras CRIS selecciona la muestra más relevante de cada grupo. Sin embargo, en este caso extremo las soluciones alcanzadas por ambas técnicas es la misma. Note además, que aunque RIS no tiene la necesidad de hacer *clustering* sobre los datos, se puede observar que RIS selecciona una muestra de cada *cluster*. En resumen, en este caso hipotético se puede observar que las dos metodologías propuestas preservan la estructura global de la clase mayoritaria encontrando los puntos que minimizan la divergencia entre su PDF y la PDF de la clase mayoritaria original.

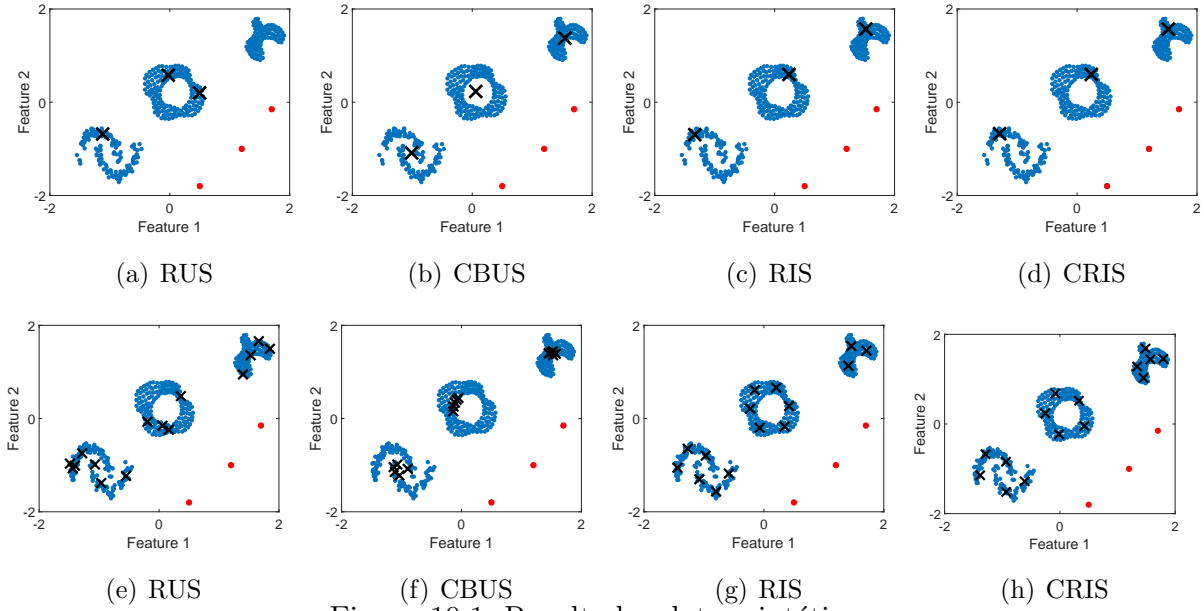


Figura 10.1: Resultados datos sintéticos

En la segunda fila se puede observar los mismos métodos, pero en este caso se pretende seleccionar 15 muestras de la clase mayoritaria para balancear al orden 5:1. Como en el caso anterior, RUS 10.1(e) selecciona puntos aleatoriamente sin consideración de la estructura de los datos y de nuevo la forma general de los datos es ignorada. También, CBUS 10.1(f) en este caso, selecciona las 5 muestras más cercanas al centroide de cada *cluster*, pero como en el anterior experimento, estas muestras no representan al resto del grupo debido a la particular forma de los datos. Por otra parte, en este segundo experimento RIS y CRIS difieren entre sus soluciones. Como en el primer experimento, RIS selecciona las 15 muestras más relevantes entre todas las muestras pertenecientes a la clase mayoritaria, por lo tanto, en la figura 10.1 (g) hay algunos *clusters* con diferentes números de muestras, dando prioridad al *cluster* central y colocando solo tres muestras en el *cluster* superior. Por el contrario, CRIS 10.1(h)

selecciona las 5 muestras más relevantes de cada grupo, obteniendo el mismo número de muestras para cada conglomeración de datos, lo cual para propósitos de clasificación podría convertirse en una ventaja, ya que esto garantiza que para cada estructura de datos el mismo número de puntos serán seleccionados.

10.2. Resultados de las pruebas sobre las bases de datos del repositorio KEEL

Como se especificó en el capítulo anterior, se realizaron pruebas sobre 44 bases de datos de pequeña escala. Concretamente, se submuestra la clase mayoritaria de cada una de las bases de datos con el método RIS y posteriormente cuatro clasificadores son entrenados para encontrar la mejor combinación de métodos. Los clasificadores empleados son LDA, KNN, SVM y MLP. En la tabla 10.1 se pueden ver los resultados obtenidos de cada clasificador en términos del AUC para cada una de las bases de datos. Para determinar cuál es el clasificador más efectivo se realiza una prueba de *t-student* con un nivel de significancia $\alpha = 0,05$. La prueba se realiza para cada una de las bases de datos y por cada pareja de clasificadores y se cuenta el número de veces que cada uno de los métodos fue estadísticamente superior a los demás. El resultado de este procedimiento evidencia que la SVM es el clasificador que mejor rendimiento presenta, siendo superior estadísticamente en 13 bases de datos, seguido del LDA con 6 victorias, y en último lugar MLP y KNN, ambos ganadores en 4 bases de datos.

De acuerdo a lo anterior, se escoge la SVM como el clasificador más apropiado. También, para el método CRIS se escogió la SVM como el clasificador para desarrollar el ensamble. Sin embargo, es necesario ajustar el número M de clasificadores a combinar (ver sección 8). Para esto se prueba combinando tres, cuatro y cinco clasificadores. Los resultados se muestran en la tabla 10.2, evidenciando que el uso de tres clasificadores (CRIS3) gana en 15 de las 44 bases de datos contra 11 victorias del CRIS4 y 9 del CRIS5. Además, el uso de tres clasificadores fue la combinación con el mejor promedio en el rendimiento de clasificación en términos de la AUC.

Finalmente, se comparan los resultados obtenidos por los métodos propuestos contra los métodos que mejor rendimiento han mostrado en el estado del arte para estas bases de datos. En consecuencia, en la tabla 10.3 se muestra la comparación de los resultados de los métodos anteriormente descritos y se resalta en negrita el mejor método. Según lo indicado por los resultados de clasificación, los enfoques propuestos de CRIS y RIS con SVM demostraron el rendimiento de clasificación más alto y segundo más alto con respecto al AUC promedio. Para los conjuntos de datos de pequeña escala, en comparación con los cuatro enfoques del estado del arte, el método RIS gana en el 56,8% de los casos mientras que con la técnica

Tabla 10.1: Rendimiento de clasificación de las diferentes clasificadores para las bases de datos de prueba

Base de datos	Classificadores			
	LDA	k-NN	SVM	MLP
Abalone9-18	0,780±0,403	0,715±0,298	0,761±0,365	0,767±0,378
Abalone19	0,843±0,069	0,812±0,071	0,812±0,071	0,833±0,083
Ecoli-0_vs_1	0,992±0,013	0,977±0,031	0,984±0,018	0,99±0,012
Ecoli-0-1-3-7_vs_2-6	0,882±0,081	0,78±0,437	0,836±0,237	0,898±0,131
Ecoli1	0,952±0,025	0,944±0,038	0,932±0,045	0,933±0,036
Ecoli2	0,918±0,051	0,941±0,06	0,956±0,029	0,901±0,058
Ecoli3	0,924±0,037	0,939±0,045	0,939±0,021	0,915±0,048
Ecoli4	0,995±0,006	0,994±0,003	0,984±0,022	0,941±0,106
Glass0	0,743±0,052	0,819±0,064	0,879±0,056	0,78±0,099
Glass0123vs456	0,931±0,034	0,927±0,015	0,955±0,025	0,955±0,026
Glass016vs2	0,652±0,131	0,643±0,129	0,795±0,153	0,6576±0,125
Glass016vs5	0,886±0,239	0,876±0,07	0,906±0,074	0,974±0,044
Glass1	0,619±0,15	0,791±0,062	0,712±0,062	0,701±0,071
Glass2	0,671±0,239	0,602±0,202	0,859±0,099	0,73±0,173
Glass4	0,912±0,065	0,914±0,06	0,937±0,066	0,911±0,082
Glass5	0,829±0,133	0,833±0,17	0,909±0,097	0,971±0,04
Glass6	0,947±0,045	0,94±0,053	0,974±0,031	0,905±0,134
Haberman	0,685±0,056	0,642±0,052	0,693±0,021	0,703±0,076
Iris0	1±0	1±0	1±0	1±0
New-thyroid1	0,999±0,002	0,995±0,008	0,987±0,03	0,994±0,012
New-thyroid2	0,996±0,005	0,994±0,009	0,983±0,025	0,998±0,002
Page-blocks0	0,908±0,028	0,964±0,006	0,974±0,004	0,972±0,006
Page-blocks13vs4	0,887±0,041	0,974±0,003	0,991±0,013	0,965±0,014
Pima	0,829±0,025	0,79±0,022	0,815±0,027	0,809±0,047
Segmemt0	0,986±0,013	0,985±0,011	0,999±0,001	0,977±0,02
Shuttle0vs4	0,996±0,009	1±0	1±0	1±0
Shuttle2vs4	0,809±0,185	1±0	1±0	0,992±0,018
Vehicle0	0,989±0,007	0,961±0,019	0,99±0,005	0,987±0,007
Vehicle1	0,861±0,032	0,798±0,016	0,907±0,027	0,842±0,037
Vehicle2	0,991±0,006	0,981±0,009	0,993±0,005	0,988±0,009
Vehicle3	0,844±0,028	0,793±0,017	0,874±0,015	0,85±0,035
Vowel0	0,982±0,0101	0,987±0,0064	0,999±0,001	0,997±0,004
Wisconsin	0,995±0,004	0,976±0,008	0,98±0,005	0,991±0,004
Yeast05679vs4	0,831±0,049	0,849±0,051	0,833±0,054	0,842±0,052
Yeast1	0,783±0,023	0,787±0,027	0,777±0,026	0,758±0,048
Yeast1vs7	0,835±0,049	0,809±0,08	0,808±0,06	0,845±0,06
Yeast1289vs7	0,751±0,096	0,743±0,135	0,638±0,126	0,769±0,146
Yeast1458vs7	0,65±0,113	0,657±0,135	0,592±0,183	0,634±0,123
Yeast2vs4	0,926±0,018	0,966±0,025	0,938±0,039	0,943±0,038
Yeast2vs8	0,765±0,127	0,804±0,144	0,753±0,118	0,753±0,12
Yeast3	0,965±0,008	0,958±0,019	0,961±0,006	0,962±0,008
Yeast4	0,862±0,026	0,902±0,04	0,859±0,079	0,889±0,019
Yeast5	0,985±0,004	0,964±0,006	0,987±0,005	0,972±0,013
Yeast6	0,922±0,092	0,903±0,079	0,928±0,076	0,916±0,062
Promedio	0,875±0,109	0,878±0,112	0,987±0,621	0,889±0,104

CAPÍTULO 10. RESULTADOS DE LAS PRUEBAS SOBRE BASES DE DATOS DE PEQUEÑA ESCALA

Tabla 10.2: Rendimiento de clasificación de diferente número de clasificadores en el método CRIS.

Base de datos	CRIS3	CRIS4	CRIS5
Abalone9-18	0,955 0,032	0,936 0,044	0,929 0,045
Abalone19	0,824±0,079	0,801±0,096	0,770±0,132
Ecoli-0_vs_1	1,000±0	1,000±0,000	1,000±0,000
Ecoli-0-1-3-7_vs_2-6	0,946±0,077	0,912±0,127	0,855±0,291
Ecoli1	0,951±0,025	0,944±0,032	0,958±0,027
Ecoli2	0,944±0,039	0,946±0,046	0,955±0,039
Ecoli3	0,942±0,022	0,940±0,016	0,951±0,011
Ecoli4	0,985±0,021	0,981±0,020	0,986±0,019
Glass0	0,867±0,055	0,859±0,043	0,876±0,038
Glass0123vs456	0,979±0,013	0,979±0,010	0,978±0,007
Glass016vs2	0,838±0,096	0,816±0,067	0,754±0,178
Glass016vs5	0,963±0,075	0,937±0,063	0,983±0,019
Glass1	0,812±0,065	0,803±0,094	0,796±0,080
Glass2	0,815±0,150	0,849±0,085	0,850±0,136
Glass4	0,967±0,031	0,989±0,012	0,970±0,032
Glass5	0,978±0,033	0,944±0,055	0,978±0,030
Glass6	0,978±0,023	0,976±0,021	0,974±0,025
Haberman	0,603±0,069	0,613±0,080	0,612±0,083
Iris0	1,000±0,000	1,000±0,000	1,000±0,000
New-thyroid1	0,999±0,003	0,998±0,003	0,998±0,004
New-thyroid2	0,997±0,004	0,998±0,002	0,998±0,004
Page-blocks0	0,984±0,005	0,984±0,006	0,985±0,005
Page-blocks13vs4	0,995±0,005	0,996±0,005	0,994±0,005
Pima	0,805±0,045	0,826±0,019	0,812±0,020
Segment0	1,000±0,000	1,000±0,000	1,000±0,000
Shuttle0vs4	1,000±0,000	1,000±0,000	1,000±0,000
Shuttle2vs4	1,000±0,000	1,000±0,000	1,000±0,000
Vehicle0	0,998±0,002	0,997±0,001	0,996±0,003
Vehicle1	0,917±0,022	0,924±0,025	0,925±0,020
Vehicle2	0,997±0,005	0,995±0,003	0,996±0,004
Vehicle3	0,912±0,016	0,897±0,014	0,900±0,020
Vowel0	1,000±0,000	1,000±0,000	1,000±0,000
Wisconsin	0,991±0,007	0,992±0,005	0,993±0,006
Yeast05679vs4	0,870±0,046	0,868±0,041	0,867±0,056
Yeast1	0,783±0,035	0,778±0,021	0,781±0,016
Yeast1vs7	0,836±0,076	0,840±0,057	0,840±0,053
Yeast1289vs7	0,793±0,083	0,780±0,040	0,773±0,100
Yeast1458vs7	0,671±0,087	0,660±0,106	0,715±0,069
Yeast2vs4	0,979±0,031	0,975±0,014	0,959±0,017
Yeast2vs8	0,809±0,082	0,785±0,096	0,807±0,131
Yeast3	0,971±0,015	0,967±0,012	0,968±0,010
Yeast4	0,897±0,033	0,888±0,038	0,904±0,029
Yeast5	0,986±0,004	0,988±0,002	0,987±0,005
Yeast6	0,930±0,086	0,935±0,069	0,927±0,902
Promedio	0,920±0,037	0,916±0,032	0,916±

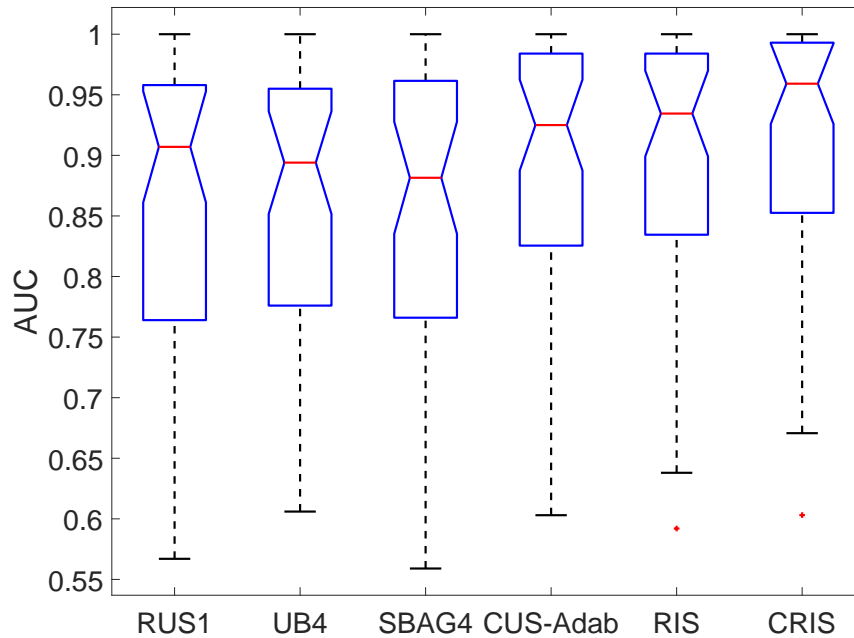


Figura 10.2: Boxplot de los resultados de clasificación

CRIS, los resultados mejoran, ganando en el 70,45 % de los casos. Es importante destacar que aunque el enfoque RIS es solo un método de submuestreo y sus resultados de clasificación se lograron al entrenar un clasificador, fue superior a RUS1, UB4, SBAG4 y CBUS-AB, que son métodos conjuntos que entrenan un mayor número de clasificadores.

Además, se han resumido los resultados para las bases de datos de pequeña escala de los métodos de comparación en la figura 10.2 utilizando el diagrama de *box plot* como esquema de representación. Los *box plot* son una herramienta para el reporte de datos que permite la representación gráfica del rendimiento de los algoritmos, indicando características esenciales como la mediana, valores extremos y difusión de valores sobre la mediana en forma de cuartiles [38].

Se pueda observar que los *box plot* de CRIS y RIS son los más compactos; ambos métodos tienen mejores resultados que el resto, pero el valor de la mediana de CRIS es el mejor. Aunque CRIS se basa en la técnica de submuestreo basada en agrupamiento, nuestra versión mejora notablemente los resultados logrados por CBUS-AdaBoost al pasar de un valor en la mediana de 0,925 a 0,959.

CAPÍTULO 10. RESULTADOS DE LAS PRUEBAS SOBRE BASES DE DATOS DE PEQUEÑA ESCALA

Tabla 10.3: Rendimiento de clasificación de las diferentes técnicas para las bases de datos de prueba

Base de datos	OD	Métodos del estado del arte				nuestros métodos	
		RUS1 [36]	UB4 [39]	SBAG4 [37]	CBUS-AB [9]	RIS	CRIS
Abalone9-18	16,68	0,693	0,719	0,745	0,831	0,893	0,955
Abalone19	128,87	0,631	0,721	0,572	0,728	0,812	0,824
Ecoli-0_vs_1	1,86	0,969	0,98	0,983	0,982	0,984	1,000
Ecoli-0-1-3-7_vs_2-6	39,15	0,794	0,745	0,828	0,804	0,836	0,946
Ecoli1	3,36	0,883	0,9	0,9	0,927	0,932	0,951
Ecoli2	5,46	0,899	0,884	0,888	0,947	0,956	0,944
Ecoli3	8,19	0,856	0,908	0,885	0,926	0,939	0,942
Ecoli4	13,84	0,942	0,888	0,933	0,95	0,984	0,985
Glass0	3,19	0,813	0,814	0,839	0,873	0,879	0,867
Glass0123vs456	10,29	0,93	0,904	0,946	0,97	0,965	0,979
Glass016vs2	19,44	0,617	0,754	0,559	0,79	0,795	0,838
Glass016vs5	1,82	0,989	0,943	0,866	0,964	0,906	0,963
Glass1	10,39	0,763	0,737	0,728	0,824	0,712	0,812
Glass2	15,47	0,78	0,769	0,779	0,76	0,859	0,815
Glass4	22,81	0,915	0,846	0,874	0,853	0,937	0,967
Glass5	22,81	0,943	0,949	0,878	0,949	0,909	0,978
Glass6	6,38	0,918	0,904	0,931	0,905	0,974	0,978
Haberman	2,68	0,655	0,664	0,656	0,603	0,693	0,603
Iris0	2	0,99	0,99	0,98	0,99	1,000	1,000
New-thyroid1	5,14	0,958	0,964	0,975	0,973	0,987	0,999
New-thyroid2	4,92	0,938	0,958	0,961	0,924	0,983	0,997
Page-blocks0	8,77	0,948	0,958	0,953	0,986	0,974	0,984
Page-blocks13vs4	15,85	0,987	0,978	0,988	0,992	0,991	0,995
Pima	1,9	0,726	0,76	0,751	0,758	0,815	0,805
Segment0	6,01	0,993	0,988	0,994	0,996	0,999	1,000
Shuttle0vs4	13,87	1	1	1	1	1,000	1,000
Shuttle2vs4	20,5	1	1	1	0,988	1,000	1,000
Vehicle0	3,23	0,958	0,952	0,965	0,99	0,990	0,998
Vehicle1	2,52	0,747	0,787	0,769	0,832	0,907	0,917
Vehicle2	2,52	0,97	0,964	0,966	0,995	0,993	0,997
Vehicle3	2,52	0,765	0,802	0,763	0,827	0,874	0,912
Vowel0	10,1	0,943	0,947	0,988	0,987	0,999	1,000
Wisconsin	1,86	0,964	0,96	0,96	0,99	0,980	0,991
Yeast05679vs4	9,35	0,803	0,794	0,818	0,869	0,833	0,870
Yeast1	2,46	0,719	0,722	0,734	0,747	0,777	0,783
Yeast1vs7	13,87	0,715	0,786	0,697	0,768	0,808	0,836
Yeast1289vs7	30,56	0,721	0,734	0,658	0,692	0,638	0,793
Yeast1458vs7	22,1	0,567	0,606	0,623	0,627	0,592	0,671
Yeast2vs4	9,08	0,933	0,936	0,897	0,977	0,938	0,979
Yeast2vs8	23,1	0,789	0,783	0,784	0,868	0,753	0,809
Yeast3	8,11	0,925	0,934	0,944	0,967	0,961	0,971
Yeast4	28,41	0,812	0,855	0,773	0,874	0,859	0,897
Yeast5	32,78	0,959	0,952	0,962	0,987	0,987	0,986
Yeast6	39,15	0,823	0,869	0,836	0,909	0,928	0,930
Promedio		0,856	0,866	0,853	0,889	0,900	0,922
Victorias		1/44	0/44	0/44	4/44	7/44	32/44

Capítulo 11

Resultados de las pruebas sobre las bases de datos de gran escala

A continuación, se presentan los resultados de todas las pruebas realizadas, sobre cada una de las tres bases de datos de gran escala mencionadas en la sección [9.1.1](#).

11.1. Resultados bases de datos de cáncer y proteínas

Las figuras [11.1](#) y [11.2](#) muestran los resultados de clasificación obtenidos con algunos métodos de comparación reportados en el estado del arte. Como se indica en los resultados de clasificación, para ambas bases de datos, el método CRIS demostró el mayor rendimiento en términos del AUC. Sin embargo, el método RIS alcanzó un rendimiento similar pero inferior comparado con el método CBUS-Adaboost. Lo anterior puede ser explicado debido a que RIS selecciona más puntos en las regiones donde la PDF estimada de la clase mayoritaria tiene los valores más grandes. Esta propiedad es altamente problema-dependiente, pero para bases de datos de gran escala, el método RIS se encuentra en desventaja ya que elige mayormente puntos de las zonas más densas. De acuerdo con esto, CRIS y CBUS-Adaboost tienen la ventaja, ya que al agrupar la bases de datos en $k = N_-$ grupos, permiten seleccionar puntos en todas las regiones del espacio de características. Sin embargo, CBUS selecciona el punto más cercano al centroide de cada grupo, lo cual es un gran problema porque los valores atípicos o formas no Gaussianas pueden hacer que el centro del centroide carezca de significado para representar el grupo. Por otra parte, la contribución más significativa del método propuesto, es que CRIS usa todos los estadísticos de más alto orden expresados por el kernel, para extraer mayor información de cada conglomerado y asignar los puntos adecuados que más se ajustan a las propiedades estructurales de cada grupo.

CAPÍTULO 11. RESULTADOS DE LAS PRUEBAS SOBRE LAS BASES DE DATOS DE GRAN ESCALA

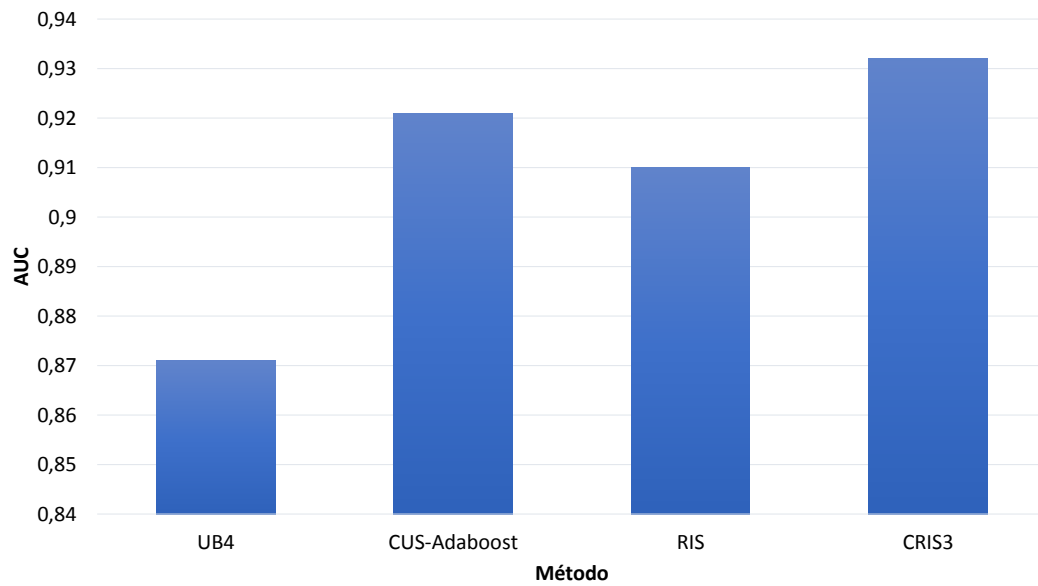


Figura 11.1: Resultados de clasificación para la base de datos de cáncer.

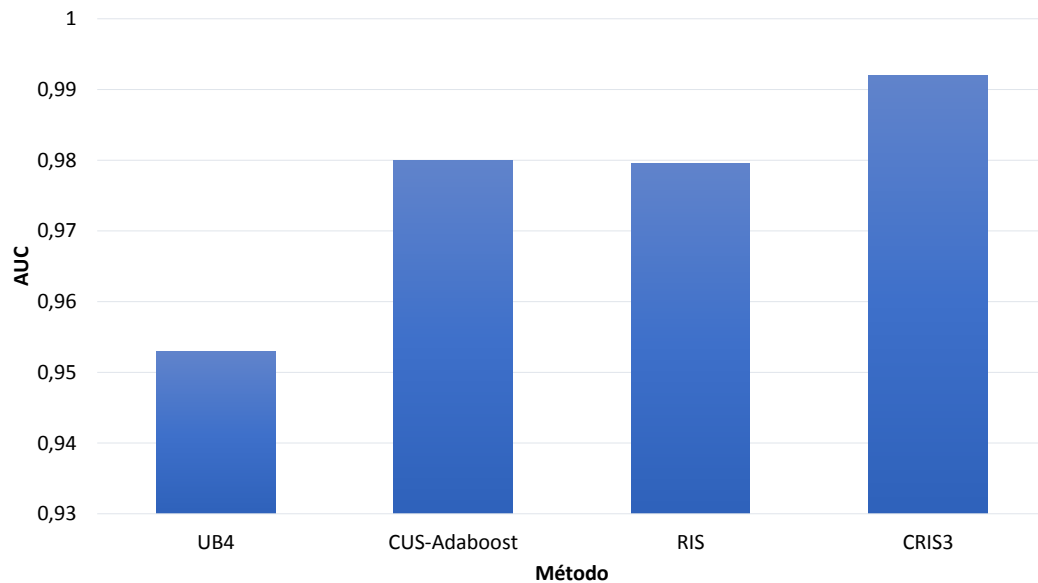


Figura 11.2: Resultados de clasificación para la base de datos de proteínas.

11.2. Resultados base de datos de FCD

La tabla 11.1 muestra los resultados obtenidos para los métodos propuestos; WUS, RUS, *Bagging*, CBUS versión 1 y la combinación de CBUS y *Bagging*. Sin submuestreo el clasificador alcanza una alta especificidad (99.7%), pero la sensibilidad es baja (49.8%). Esto sucede porque el clasificador está sesgado hacia la clase sana debido al alto desbalance entre las clases. Cuando los datos de entrada son aleatoriamente submuestreados, los resultados de sensibilidad mejoran, pero la especificidad disminuye. Sin embargo, el valor de la media geométrica (*G-mean*), que abarca información de la sensibilidad y la especificidad, aumenta significativamente. El método de *Bagging* mejora el valor del *G-mean*, con respecto a los dos métodos previos. Para CBUS este valor es aún mayor, debido a que esta técnica selecciona las muestras más relevantes de grupos de datos donde se concentran los datos. La combinación de CBUS y *Bagging* mejora los resultados de clasificación. La figura 11.3, presenta las curvas ROC para los métodos propuestos, donde se evidencia que el método propuesto se desempeñó mejor que los métodos restantes.

Método	Desbalance	media-G (%)	Sensibilidad(%)	Especificidad(%)
WUS	42:1	70.46 \pm 1.34	49.8 \pm 1.22	99.7 \pm 1.28
RUS	1:1	86.19 \pm 2.28	84.1 \pm 2.36	88.4 \pm 1.82
Bagging	5:1	89.46 \pm 0.71	88.95 \pm 0.78	89.98 \pm 0.85
CBUS	1:1	92.19 \pm 1.18	92.6 \pm 1.92	91.8 \pm 2.17
CBUS-Bagging	5:1	94.15 \pm 0.21	95.11 \pm 0.49	93.2 \pm 0.33

Tabla 11.1: Comparación de resultados

Uno de las investigaciones principales en esta base de datos es el estudio desarrollado por [33], el cual no considera el desbalance de las clases, sin embargo, posprocesan los resultados de salida del clasificador y agrupan los voxels de la clase enferma conectados y descartan los grupos más pequeños. Su método de detección de FCD se considera exitoso si este grupo se superpone a la etiqueta marcada por los expertos. Bajo este criterio, su método fue capaz de detectar 16 de 22 FCD (73%). Desafortunadamente, no se contó con las resonancias magnéticas originales, por lo que no fue posible validar utilizando una estrategia análoga. Sin embargo, los resultados al nivel de observaciones individuales demuestran que los métodos propuestos superan los enfoques del estado del arte para la detección automática de FCD.

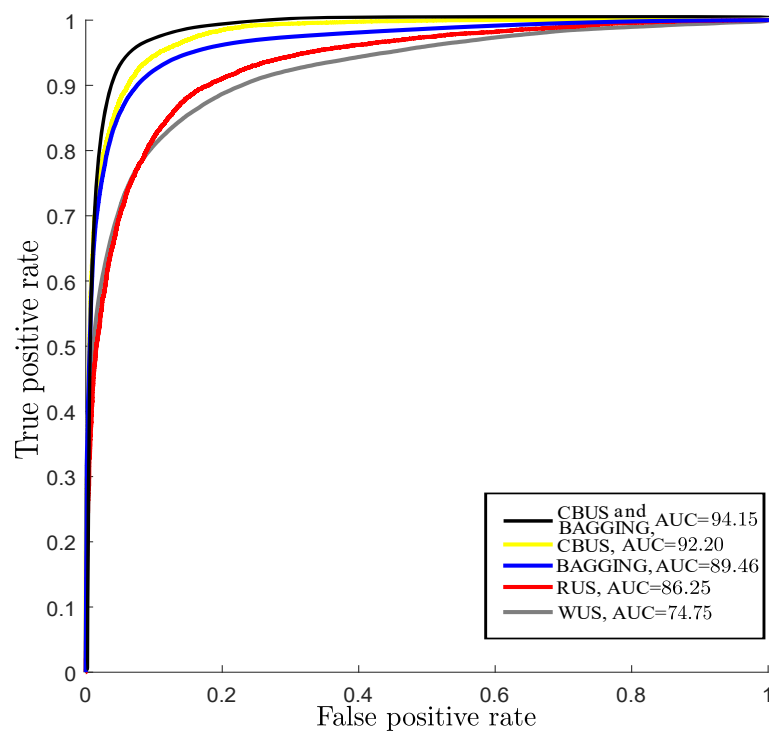


Figura 11.3: Curvas ROC de los diferentes métodos en estudio.

Parte IV

Conclusiones

Capítulo 12

Conclusiones

En este documento, se presenta una metodología de clasificación de datos desbalanceados a partir de métodos de submuestreo y métodos de ensamble. En consecuencia, se genera un esquema metodológico, con aplicabilidad a múltiples propósitos relacionados con el reconocimiento automático de patrones inusuales. Inicialmente, se hace uso del submuestreo basado en agrupamiento con el fin de codificar la estructura global de los datos. Este método mostró ser adecuado en bases de datos con grandes cantidades de datos, y donde el nivel de desbalance era muy alto en comparación con otros conjuntos de datos. Por otra parte, se propone una novedosa técnica de submuestreo basada en un principio de auto-organización de información relevante, llamado *submuestreo basado en información relevante* (RIS), para apoyar la clasificación supervisada en el contexto de clases desbalanceadas. Con este fin, el RIS encuentra las M muestras con entropía mínima al tiempo que conserva diferentes niveles de detalle sobre la clase mayoritaria al minimizar también su divergencia Cauchy-Schwarz. Además, la técnica RIS calcula medidas en el marco de la teoría de la información que cuantifican la microestructura estadística de la clase mayoritaria más allá de las estadísticas de segundo orden, lo que hace que el método sea sólido para las distribuciones de datos complejos. Además, proponemos un enfoque híbrido entre el submuestreo basado en agrupamiento (CBUS), el método RIS y el método de ensamble de *Bagging*. La nueva variación, llamada CRIS, produce k grupos sobre la clase mayoritaria mediante el algoritmo k -means. Aquí, el número de grupos k se establece para que sea igual al número de muestras de datos en la clase minoritaria ($k = N_-$). Luego, se ejecuta el algoritmo RIS pero seleccionando tantas muestras en cada grupo como los clasificadores que se desee combinar. Este método tiene como fin explotar las ventajas de cada uno de los métodos combinados, ya que CBUS mostró ser una buena alternativa en presencia de grandes cantidades de datos y también ayuda a preservar la estructura global de los datos, mientras que RIS codifica apropiadamente la estructura local e interna de un conjunto de datos.

Los métodos fueron probados en 44 conjuntos de datos de clasificación binaria de baja escala, proveniente del repositorio de conjuntos de datos de KEEL. También, se probaron las diferentes metodologías en dos conjuntos de datos a gran escala de Knowledge Discovery y Data Mining Cup. Los resultados alcanzados en este trabajo muestran que, efectivamente, la metodología propuesta mejora el desempeño de clasificación de las bases de datos desbalanceadas, reduciendo la pérdida de información y el sobre ajuste, comparado con los métodos tradicionales.

Como trabajo futuro, se pretende hacer aproximaciones de la matriz *kernel* que permitan usar el método en grandes cantidades de datos sin hacer uso de un ensamble. Se desea también formular un enfoque de submuestreo que considere la clase minoritaria dentro de la función de costo, lo cual podría disminuir la superposición de clases. Además, se busca manipular la función de costo para ponderar los puntos de la clase mayoritaria de acuerdo con las regiones donde se ubican para evitar que muchos puntos se coloquen en áreas más densas, lo que reduciría la redundancia.

Apéndice A

Teoría de la información

El aprendizaje basado en teoría de la información (*Information theoretic learning*-ITL) se basa en la entropía de Renyi de orden α [40]. En particular, sea X una variable aleatorio con *función de densidad de probabilidad*-(PDF) $f(x)$, donde $x \in \mathcal{X}$ y \mathcal{X} es un conjunto finito, la entropía- α $H_\alpha(X)$ está definida como:

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \int_{\mathcal{X}} f^\alpha(x) dx. \quad (\text{A.1})$$

Notablemente, el caso límite cuando $\alpha \rightarrow 1$ conlleva a la entropía de Shannon ([41]). también, el caso de $\alpha = 2$ es de particular interés porque este provee un estimador de una forma muy simple. De hecho, la entropía cuadrática de Renyi, $H_2(X) = -\log \int_{\mathcal{X}} f^2(x) dx$, puede ser estimada directamente a partir de un conjunto de muestras independientes e idénticamente distribuidas (i.i.d) $\{x_i \in \mathcal{X}\}_{i=1}^N$ usando una función basada en *kernel* $\kappa_\sigma : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ para aproximar la PDF de la siguiente manera:

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^N \kappa_\sigma(x, x_i); \quad (\text{A.2})$$

así, asumiendo *kernel* Gaussianos $G_\sigma(\cdot)$ con parámetro de escala σ y sustituyéndolo en la expresión de la entropía cuadrática de Renyi, se obtiene su estimador:

$$\begin{aligned}\hat{H}_2(X) &= -\log \int_{\mathcal{X}} \left(\frac{1}{N} \sum_{i=1}^N G_{\sigma}(x, x_i) \right)^2 dx \\ &= -\log \left(\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N G_{\sqrt{2}\sigma}(x_i, x_j) \right),\end{aligned}\tag{A.3}$$

donde el argumento del logaritmo en la entropía de Renyi cuadrática se conoce como *Potencial de información*(IP) y se puede estimar directamente a partir de los datos como:

$$V(X) = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N G_{\sqrt{2}\sigma}(x_i, x_j).\tag{A.4}$$

Por otra parte, en el marco del ITL, hay una definición de “distancia.entre PDF, que es compatible con la entropía de Renyi de orden 2: la divergencia *Cauchy-Schwarz* (CS). Ahora, sean $\{x_i \in \mathcal{X}\}_{i=1}^N$ y $\{y_i \in \mathcal{Y}\}_{i=1}^M$ muestras i.i.d provenientes de dos variables aleatorias X y Y con PDF $f(x)$ y $g(y)$ respectivamente. Luego, la divergencia CS entre X y Y es dada por:

$$\begin{aligned}D_{CS}(f, g) &= -\log \left(\int_{\mathcal{Y}} \int_{\mathcal{X}} f(x)g(y)dx dy \right)^2 \\ &\quad + \log \int_{\mathcal{X}} f^2(x)dx + \log \int_{\mathcal{Y}} g^2(y)dy \\ &= 2H_2(X, Y) - H_2(X) - H_2(Y),\end{aligned}\tag{A.5}$$

donde el primer término es la *cross-entropía cuadrática de Renyi*. Similarmente, se puede estimar $H_2(X, Y)$ como $\hat{H}_2(X, Y) = -\log V(X, Y)$, donde $V(X, Y)$ es el potencial de información cruzado y de nuevo, este toma una forma simple para ser estimado a partir de los datos de la siguiente manera:

$$V(X, Y) = \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M G_{\sqrt{2}\sigma}(x_i, y_j),\tag{A.6}$$

Ya que los *kernels* son funciones positivas que decrecen a medida que la distancia entre las muestras aumenta, existe una analogía física en la estimación de la entropía de Renyi. Se puede pensar que cada observación crea un “campo de potencial de información.”^{en} el espacio de las

muestras, tal como las partículas físicas crean un campo gravitacional. En este contexto, cada instancia puede ser nombrada como una partícula de información e interactúan en el campo de potencial de información creando fuerzas de información [42].

Para estimar la fuerza de información ejercida sobre la muestra \mathbf{x}_k debida a las otras muestras, se debe calcular la derivada del potencial de información respecto a la posición de la muestra. Para kernel Gaussianos, esta derivada es fácilmente evaluada como:

$$F(\mathbf{x}_k, \mathbf{X}) = \frac{\partial V(\mathbf{X})}{\partial \mathbf{x}_k} = \frac{1}{2N\sigma^2} \sum_{j=1}^N G_{\sqrt{2}\sigma}(\mathbf{x}_k, \mathbf{x}_j)(\mathbf{x}_k - \mathbf{x}_j), \quad (\text{A.7})$$

por lo tanto, según esta ecuación, las partículas más lejanas a cada muestra ejercen menos fuerza que aquellas que se encuentran más cercanas.

Bibliografía

- [1] M. Maalouf and T. B. Trafalis, “Rare events and imbalanced datasets: an overview,” *International Journal of Data Mining, Modelling and Management*, vol. 3, no. 4, pp. 375–388, 2011.
- [2] N. Japkowicz and S. Stephen, “The class imbalance problem: A systematic study,” *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.
- [3] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.
- [4] R. Akbani, S. Kwek, and N. Japkowicz, “Applying support vector machines to imbalanced datasets,” *Machine learning: ECML 2004*, pp. 39–50, 2004.
- [5] V. López, S. del Río, J. M. Benítez, and F. Herrera, “Cost-sensitive linguistic fuzzy rule based classification systems under the mapreduce framework for imbalanced big data,” *Fuzzy Sets and Systems*, vol. 258, pp. 5–38, 2015.
- [6] S.-J. Yen and Y.-S. Lee, “Cluster-based under-sampling approaches for imbalanced data distributions,” *Expert Systems with Applications*, vol. 36, no. 3, pp. 5718–5727, 2009.
- [7] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, “An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics,” *Information Sciences*, vol. 250, pp. 113–141, 2013.
- [8] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, “A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, 2012.
- [9] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, “Clustering-based undersampling in class-imbalanced data,” *Information Sciences*, vol. 409, pp. 17–26, 2017.

- [10] J. Prusa, T. M. Khoshgoftaar, D. J. Dittman, and A. Napolitano, "Using random undersampling to alleviate class imbalance on tweet sentiment data," in *Information Reuse and Integration (IRI), 2015 IEEE International Conference on*. IEEE, 2015, pp. 197–202.
- [11] G. Di Leo, C. Liguori, A. Pietrosanto, and P. Sommella, "A vision system for the online quality monitoring of industrial manufacturing," *Optics and Lasers in Engineering*, vol. 89, pp. 162–168, 2017.
- [12] A. Singla and A. Sharma, "Physical access system security of iot devices using machine learning techniques," *Available at SSRN 3356785*, 2019.
- [13] J. Shan, S. K. Alam, B. Garra, Y. Zhang, and T. Ahmed, "Computer-aided diagnosis for breast ultrasound using computerized bi-rads features and machine learning methods," *Ultrasound in medicine & biology*, vol. 42, no. 4, pp. 980–988, 2016.
- [14] D. Avci and A. Dogantekin, "An expert diagnosis system for parkinson disease based on genetic algorithm-wavelet kernel-extreme learning machine," *Parkinson's disease*, vol. 2016, 2016.
- [15] K. Matsuda and K. Murase, "Single-layered complex-valued neural network with smote for imbalanced data classification," in *Soft Computing and Intelligent Systems (SCIS) and 17th International Symposium on Advanced Intelligent Systems, 2016 Joint 8th International Conference on*. IEEE, 2016, pp. 349–354.
- [16] N. Chawla, N. Japkowicz, and A. Kolcz, "Workshop learning from imbalanced data sets ii," in *Proc. Int Conf. Machine Learning*, 2003.
- [17] M. Tahir, J. Kittler, K. Mikolajczyk, and F. Yan, "A multiple expert approach to the class imbalance problem using inverse random under sampling," *Multiple Classifier Systems*, pp. 82–91, 2009.
- [18] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou, "A novel ensemble method for classifying imbalanced data," *Pattern Recognition*, vol. 48, no. 5, pp. 1623–1637, 2015.
- [19] A. D Addabbo and R. Maglietta, "Parallel selective sampling method for imbalanced and large data classification," *Pattern Recognition Letters*, vol. 62, pp. 61–67, 2015.
- [20] K. M. Ting, "An instance-weighting method to induce cost-sensitive trees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 3, pp. 659–665, 2002.
- [21] C. Zhang, W. Gao, J. Song, and J. Jiang, "An imbalanced data classification algorithm of improved autoencoder neural network," in *Advanced Computational Intelligence (ICACI), 2016 Eighth International Conference on*. IEEE, 2016, pp. 95–99.

- [22] B. Wang and J. Pineau, "Online bagging and boosting for imbalanced data streams," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 12, pp. 3353–3366, 2016.
- [23] D. Wu, Z. Wang, Y. Chen, and H. Zhao, "Mixed-kernel based weighted extreme learning machine for inertial sensor based human activity recognition with imbalanced dataset," *Neurocomputing*, vol. 190, pp. 35–49, 2016.
- [24] C. L. Castro and A. P. Braga, "Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data," *IEEE transactions on neural networks and learning systems*, vol. 24, no. 6, pp. 888–899, 2013.
- [25] B. Krawczyk and G. Schaefer, "An improved ensemble approach for imbalanced classification problems," in *Applied Computational Intelligence and Informatics (SACI), 2013 IEEE 8th International Symposium on*. IEEE, 2013, pp. 423–426.
- [26] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017.
- [27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [28] B. Wang and N. Japkowicz, "Imbalanced data set learning with synthetic samples," in *Proc. IRIS Machine Learning Workshop*, vol. 19, 2004.
- [29] H. He and Y. Ma, *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons, 2013.
- [30] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [31] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.
- [32] J. C. Principe, D. Xu, and J. Fisher, "Information theoretic learning," *Unsupervised adaptive filtering*, vol. 1, pp. 265–319, 2000.
- [33] S. Adler, K. Wagstyl, R. Gunny, L. Ronan, D. Carmichael, J. H. Cross, P. C. Fletcher, and T. Baldeweg, "Novel surface features for automated detection of focal cortical dysplasias in paediatric epilepsy," *NeuroImage: Clinical*, vol. 14, pp. 18–27, 2017.
- [34] S.-J. Hong, H. Kim, D. Schrader, N. Bernasconi, B. C. Bernhardt, and A. Bernasconi, "Automated detection of cortical dysplasia type ii in mri-negative epilepsy," *Neurology*, vol. 83, no. 1, pp. 48–55, 2014.

- [35] B. Ahmed, C. E. Brodley, K. E. Blackmon, R. Kuzniecky, G. Barash, C. Carlson, B. T. Quinn, W. Doyle, J. French, O. Devinsky *et al.*, “Cortical feature analysis and machine learning improves detection of “mri-negative” focal cortical dysplasia,” *Epilepsy & Behavior*, vol. 48, pp. 21–28, 2015.
- [36] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, “Rusboost: A hybrid approach to alleviating class imbalance,” *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 1, pp. 185–197, 2010.
- [37] S. Wang and X. Yao, “Diversity analysis on imbalanced data sets by using ensemble models,” in *Computational Intelligence and Data Mining, 2009. CIDM’09. IEEE Symposium on*. IEEE, 2009, pp. 324–331.
- [38] C. Elkan, “The foundations of cost-sensitive learning,” in *International joint conference on artificial intelligence*, vol. 17, no. 1. Lawrence Erlbaum Associates Ltd, 2001, pp. 973–978.
- [39] R. Barandela, R. M. Valdovinos, and J. S. Sánchez, “New applications of ensembles of classifiers,” *Pattern Analysis & Applications*, vol. 6, no. 3, pp. 245–256, 2003.
- [40] G. E. Crooks, “On measures of entropy and information,” *Tech. Note*, vol. 9, p. v4, 2017.
- [41] S. Yu, L. G. S. Giraldo, R. Jenssen, and J. C. Principe, “Multivariate extension of matrix-based renyi’s $\{\alpha\}$ -order entropy functional,” *arXiv preprint arXiv:1808.07912*, 2018.
- [42] D. Erdogmus, “Information theoretic learning: Renyi’s entropy and its applications to adaptive system training,” Ph.D. dissertation, University of Florida Gainesville, Florida, 2002.