GRADO EN INGENIERÍA INFORMÁTICA

CREACIÓN DE ENSEMBLES UTILIZANDO TÉCNICAS DE BAGGING PARA MEJORAR EL RENDIMIENTO DE SISTEMAS DE CLASIFICACIÓN BASADOS EN REGLAS DIFUSAS

TRABAJO DE FIN DE GRADO

Autor: Rubén Sesma Redín
Director: José Antonio Sanz Delgado
Pamplona, Marzo 2017

Índice

1. Introducción

En la denominada sociedad de la información y del conocimiento en la que vivimos, existen gran cantidad de datos que deben ser tratados y almacenados adecuadamente. Actualmente éstos se almacenan principalmente en Bases de datos y *Datawarehouses* aunque existen otros tipos de almacenes de información. En esta situación, el progreso y la innovación no se ven obstaculizados por la capacidad de almacenar y recopilar datos, sino por la capacidad de gestionar, analizar, visualizar, y descubrir conocimiento de estos datos recopilados de manera oportuna y de forma escalable.

Para resolver éstos problemas se utiliza la denominada *Minería de Datos,* que se puede definir de una manera sencilla como el estudio y tratamiento de datos masivos para extraer conclusiones e información relevante de ellos. Este campo de las ciencias de la información se centra en el descubrimiento de patrones y conocimiento a partir grandes volúmenes de información y que dicho conocimiento pueda utilizado como una herramienta más en el proceso de toma de decisiones. Hay muchas aplicaciones de este campo desde la Medicina hasta los usos fraudulentos de tarjetas.

Uno de los problemas más comunes que se tratan de resolver mediante la minería de datos son los problemas de clasificación. Estos problemas, a grandes rasgos, consisten en ser capaz de diferenciar a que clase pertenece un ejemplo en base a la información de dicho ejemplo en las variables de entrada. Un problema de clasificación consiste en predecir la clase a la pertenecerá un ejemplo en base a las características que definen a este. Para afrontar estos problemas se aprende un modelo matemático denominado *clasificador* a partir de un conjunto de ejemplos de los que se conoce su clase, es decir, están etiquetados previamente. Este conjunto de ejemplos, se denomina el conjunto de entrenamiento. Una vez aprendido el clasificador, este debe ser capaz de predecir o clasificar mediante un proceso de inferencia la clase del ejemplo que se esté considerando. Por ejemplo, si un clasificador para aplicaciones médicas ha aprendido la regla "Si fiebre es alta y tensión es alta entonces debemos recetar el medicamento X", en base a los ejemplos (pacientes) anteriores que se usaron para detectar el patrón. Esta información puede ser extra (el médico desconocía esta regla) y cuando un paciente presente esos síntomas, el sistema otorga al médico dicha respuesta.

Los *Sistemas de Clasificación Basados en Reglas Difusas (SCBRD)* son una herramienta que nos permite abordar este tipo de problemas. Consisten en un algoritmo capaz de extraer reglas interpretables por una persona, como la descrita en el párrafo anterior, y a partir de esas reglas son capaces de predecir la clase a la que pertenecerá un ejemplo a clasificar. Existen diversos algoritmos de generación de reglas, pero en este proyecto utilizaremos *FARC-HD (Fuzzy Association Rule-Based Classification model for High Dimensional problems).*

En esta introducción veremos algunas técnicas para mejorar estos sistemas, la razón por la cual decidimos abordar este tema, la descripción de los objetivos y determinaremos las pautas que seguiremos en este proyecto para intentar mejorar alguna de estas técnicas.

1.1 Ensembles

Existen técnicas para mejorar el rendimiento de los sistemas de clasificación. Una de las más habituales es la mezcla de modelos. Estos métodos se denominan *Ensemble methods* ya que combinan los resultados de varios modelos. *Bagging* y *Boosting* son dos de los más populares. En concreto Bagging es el que usaremos para su estudio en este proyecto y está basado en *Bootstrap agregation* sobre el conjunto de entrenamiento.

La clasificación en estos métodos es un proceso que tiene en cuenta el resultado de varios clasificadores. Estos modelos los denominaremos *clasificadores base* aprendidos con el mismo algoritmo, pero usando un conjunto de entrenamiento común con una peculiaridad: previamente a aprender los modelos de los clasificadores base se realiza una selección de un subconjunto de ejemplos del conjunto de entrenamiento común, de esta forma cada clasificador será entrenado con diferentes ejemplos y se aprenderán modelos de clasificación diferentes. Posteriormente, para clasificar un nuevo ejemplo se consultará a todos los clasificadores.

Un parámetro muy importante a tener en cuenta en estos modelos es el tamaño del ensemble, es decir, el número de clasificadores base que se usarán para clasificar un nuevo ejemplo. Este parámetro será objeto de estudio en este proyecto, pero existen otros parámetros a tener en cuenta en la construcción de ensembles, como puede ser el tipo de voto que se use para la clasificación, pero esta parte la describiremos en detalle en la Sección 2.4.

1.2 Motivación y objetivos

En la literatura [1] se aprecia que la creación de ensembles para problemas de clasificación ofrece mejores resultados que si aplicamos la técnica directamente sobre problemas de datos no balanceados. Sin embargo, no se ha utilizado el clasificador FARC-HD como clasificador base en ensembles para afrontar los problemas de clasificación estándar de dos o más clases.

Por este motivo, el objetivo principal de éste proyecto es la implementación de un ensemble de clasificadores FARC-HD con el objetivo de realizar un estudio experimental de esta combinación tanto para problemas de clasificación estándar como para conjuntos de datos no balanceados.

Por consiguiente, en este proyecto tratamos primero de mejorar el rendimiento de FARC-HD en problemas de clasificación estándar y no balanceados. Para ello realizaremos:

- Un análisis de la influencia del *tamaño del ensemble*: estudiaremos los resultados que nos ofrece usando 10 y 40 clasificadores de acuerdo a los valores recomendados en la literatura [1].
- Implementaremos y analizaremos la influencia del *tipo de voto*: simple y ponderado.

Además, la creación de ensembles supone generar tantos conjuntos de reglas como el tamaño que defina el ensemble, lo cual es una gran cantidad de información que podremos usar para crear un SCBRD a partir de toda la información que genera el ensemble. Se propondrá una técnica para crear un clasificador a partir de los ensembles y comprobaremos la efectividad, o no, de ésta técnica propuesta, los clasificadores obtenidos se compararán con los resultados del FARC-HD original.

2. Preliminares

2.1 Lógica difusa

La lógica difusa fue creada por Lotfi Zadeh en 1965, creada para representar matemáticamente la ambigüedad o incertidumbre. Ofrece técnicas para modelar la incertidumbre intrínseca en múltiples problemas reales. Estas técnicas permiten modelar términos lingüístico empleados en el razonamiento humano: persona alta, persona baja, persona normal… Además nos provee de un modelo interpretable puesto que se utiliza el lenguaje natural.

En la lógica clásica a la hora de determinar si un elemento pertenece a un conjunto, únicamente existen métodos para determinar si el elemento pertenece o no al conjunto {0, 1}. En cambio, el principio básico de la lógica difusa se basa en que los elementos pertenecen a un conjunto con un grado de en el intervalo [0,1], llamado grado de pertenencia.

Ejemplo: Determinar si una persona *x*, es *Alta*, en el universo *X* personas.

En la lógica clásica se determinan *Crisp sets:* $f_A(x):X \rightarrow \{0, 1\}$ donde $f_A(x)= \begin{cases} 1, & si\ x \in A \\ 0, & si\ x \notin A \end{cases}$

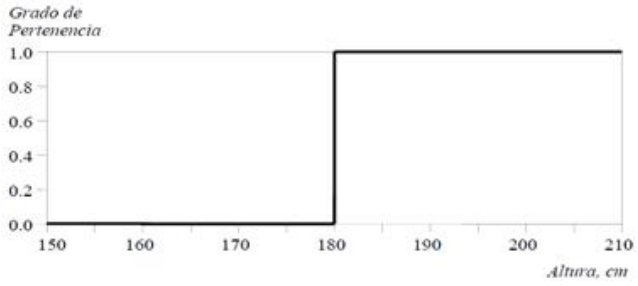

Ilustración 1Clasificación en lógica clásica

En la lógica difusa $\mu_A(x): X \rightarrow [0, 1]$ donde:

- $\mu_A(x)$:=1 si x está totalmente en A
- $\mu_A(x)$:=1 si x no está en A
- $0<\mu_A(x)<1$ si x está parcialmente en A

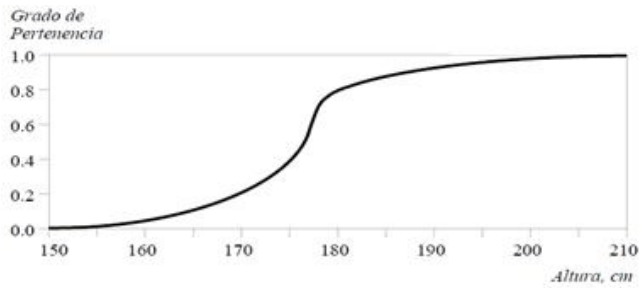

Ilustración 2Clasificación en lógica difusa

En la lógica difusa, las variables difusas se representan en base a un número de etiquetas lingüísticas y para cada etiqueta lingüística existe una función de pertenencia. Si nos fijamos en el ejemplo anterior, las funciones de pertenencia que se podrían emplear para la variable difusa altura podrían ser: Bajo, medio, alto.

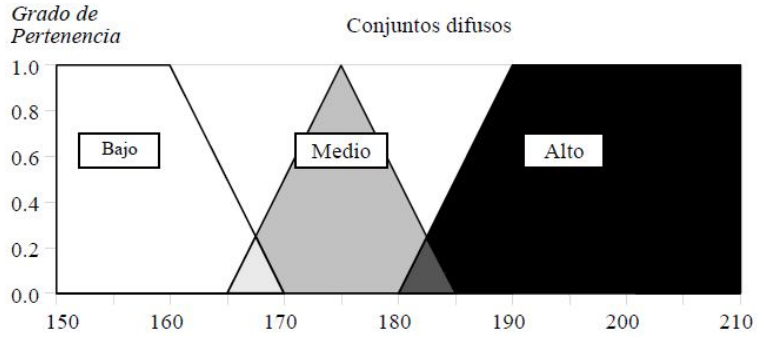

Ilustración 3 Posibles funciones de pertenencia para la variable difusa altura

Las funciones de pertenencia modelan las etiquetas lingüísticas empleadas en el problema. Estas se pueden modelar con funciones triangulares, trapezoidales o gaussianas.

En la siguiente tabla se observa la forma de clasificar si una persona es alta dependiendo de la lógica que empleemos de una forma simple: En la lógica clásica es necesario establecer un umbral, en este ejemplo se ha fijado en 180 cm, en la lógica difusa se emplean operaciones con las funciones de pertenencia para calcular el grado con el que la persona es considerada alta.

Nombre	Altura (cm)	Grado de pertenencia	
		Lógica clásica	Lógica difusa
Cristian	208	1	1.00
Marcos	205	1	1.00
Tom	198	1	0.98
Jon	181	1	0.82
David	179	0	0.78
Mikel	172	0	0.24
Steven	167	0	0.15
Francisco	158	0	0.06
Raúl	155	0	0.01
Alberto	152	0	0.00

Tabla 1 Ejemplos de clasificación

En lógica difusa, habitualmente las conjunciones de las reglas (por ejemplo, si humedad es alta y presión es muy alta) se han modelado mediante *t-normas* (aunque no son la única

técnica para modelar las conjunciones) y es lo que usamos en este proyecto. Para las disyunciones, se usan *t-conormas* y para las negaciones existen las *negaciones difusas.*

Una *t-norma* definida axiomáticamente es una función $T: [0,1]^2 \rightarrow [0,1]$ que cumple las siguientes propiedades:

i. Condiciones de límite: $T(x,1) = x,\ \forall x \in [0,1]$
ii. Monotonía: $T(x,y) \leq T(z,u)\ si\ x \leq z\ y\ y \leq u$
iii. Conmutativa: $T(x,y) = T(y,x),\ \forall x,y \in [0,1]$
iv. Asociativa: $T(T(x,y),z) = T(x,T(y,z)),\ \forall x,y,z \in [0,1]$

A pesar de existir otras, dos ejemplos de t-normas son el mínimo $T(x,y) = min\{x,y\}$ y el producto algebraico $T(x,y) = xy$.

Una *t-conorma* definida axiomáticamente es una función $S: [0,1]^2 \rightarrow [0,1]$ que cumple las siguientes propiedades:

i. Condiciones de límite: $S(x,0) = x,\ \forall x \in [0,1]$
ii. Monotonía: $S(x,y) \leq S(z,u)\ si\ x \leq z\ y\ y \leq u$
iii. Conmutativa: $S(x,y) = S(y,x),\ \forall x,y \in [0,1]$
iv. Asociativa: $S(S(x,y),z) = S(x,S(y,z)),\ \forall x,y,z \in [0,1]$

Dos ejemplos de *t-conormas* son, el máximo $S(x,y) = max\{x,y\}$y la suma algebraica $S(x,y) = x + y - xy$

Las negaciones, en lógica difusa también se modelan con funciones llamadas *negaciones difusas:*

Una función$c: [0,1] \rightarrow [0,1]$ es una *negación difusa* si cumple:

i. $c(0) = 1$ y $c(1) = 0$
ii. Monotonía: $c(x) \leq c(y),\ si\ x \geq y,\ \forall x,y \in [0,1]$

2.2 Problemas de clasificación

Clasificación es uno de los problemas más estudiados en minería de datos [2] y tiene una gran presencia en problemas reales. Un problema de clasificación consiste en una situación en la que haya que realizar la predicción de una determinada clase (atributo categórico) para un ejemplo. Para ello, la predicción se basa en los valores de los atributos/variables de entrada del ejemplo a clasificar.

El objetivo de la resolución de un problema de clasificación es construir un modelo/función, el clasificador, a partir de un conjunto de ejemplos de los que se conoce su

clase. A este conjunto de ejemplos conocidos usados para aprender el modelo, se le denomina conjunto de entrenamiento. El clasificador construido (aprendido) deberá permitir clasificar nuevos ejemplos en una de las clases existentes.

Los problemas de clasificación son problemas de aprendizaje supervisado puesto que se conoce la clase verdadera de cada uno de los ejemplos que se utilizan para construir el clasificador.

Podemos distinguir dos tipos de problemas dependiendo del número de clases que componen la salida del problema. La *clasificación estándar*, en la que los ejemplos pueden pertenecer a dos o más clases y la probabilidad de tener ejemplos de cada clase es similar. Es decir, el número de ejemplos pertenecientes a cada clase es similar. En cambio, en la clasificación con datos *no balanceados*, existen únicamente dos clases, la positiva y la negativa. Se denominan conjuntos de datos no balanceados ya que existen muchos menos ejemplos de la clase positiva (la que nos interesa clasificar bien) que de la negativa [3],[4]. Un ejemplo típico de este tipo de clasificación es la detección del cáncer. En este caso la clase positiva sería "tiene cáncer" y la negativa seria "no tiene cáncer" ya que afortunadamente existen muchísimos más pacientes sin cáncer que con cáncer y por tanto el problema no está balanceado.

2.2.1 Problemas de clasificación con datos no balanceados

En la sección anterior se ha explicado el problema que supone utilizar la tasa de acierto para evaluar el rendimiento de clasificadores sobre conjuntos de datos no balanceados, podríamos tener una buena medida pero únicamente estar acertando los ejemplos de la clase mayoritaria.

El no balanceo ó desequilibrio de las clases provocan en la mayoría de clasificadores que tiendan a tomar como ruido los casos de la clase positiva que sean muy parecidos a los de la negativa. Es decir al dominar los ejemplos de la clase negativa, aquellos ejemplos positivos con valores muy similares a los ejemplos negativos, el clasificador los interpreta como "ruido" y tiende a predecir la clase mayoritaria mucho más a menudo.

Como explica el autor en [1], en general los problemas al afrontar el aprendizaje de un clasificador para datos no balanceados son los siguientes:

- Pequeño conjunto de ejemplos positivos: Normalmente la mayoría de conjuntos no balanceados no contienen un número suficiente de ejemplos positivos como para que un algoritmo genere un buen modelo para la predicción de estos.

- Solapamiento de las clases: Como se observa en la ilustración 4a, un problema muy común en los conjuntos de datos no balanceados, además de existir pocos ejemplos positivos, también existen ejemplos negativos cuyos valores sean muy parecidos. Esto provocará que el algoritmo de aprendizaje tenga más dificultades para la generar un modelo que detecte la clase positiva.
- Pequeños conjuntos disjuntos: Esto ocurre cuando los ejemplos de la clase minoritaria se organizan en diferentes subconjuntos. Este concepto que se muestra en la ilustración 4b

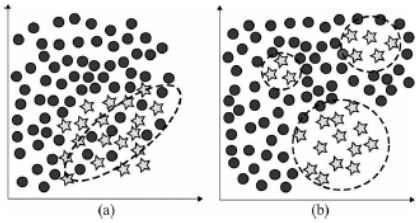

Ilustración 4 Ejemplo de problemas de los conjuntos de datos no balanceados [1]

Existen muchas técnicas para tratar de dar soluciones a estos problemas. A nivel algorítmico, se basan en modificar las características del algoritmo base asociando costes a fallar ejemplos de la clase positiva y a nivel de datos, basados en el pre-procesamiento del conjunto de entrenamiento para facilitar el aprendizaje al algoritmo.

Estas técnicas de nivel de datos cuyo objetivo es balancear el conjunto de entrenamiento, se dividen en 2 grupos de técnicas:

- *Undersampling*: Basadas en reducir instancias ruidosas de la frontera de decisión como son CNN (*Condensed nearest neighbour rule*), *Tomek links* ó NCL (*Neighbourhood cleaning rule*).
- *Oversampling*: Basadas en mantener ejemplos influyentes, añadir ejemplos de la clase positiva obtenidos mediante la interpolación de ejemplos de la clase positiva existentes, como son SMOTE (*Synthetic minority over-sampling technique*) o incluso creando prototipos como AHC (*Aglomerative Hierarchical Clustering*).

Ambas técnicas pueden combinarse y realizar un pre-procesado de los datos con técnicas híbridas.

Para evaluar la el rendimiento, o la calidad, de un clasificador, el porcentaje de acierto es una medida pero es necesario destacar que usar la tasa de acierto de un clasificador, en determinados casos, puede ser muy engañoso. Por ejemplo, en un problema donde tenemos un problema con dos clases, 9990 ejemplos pertenecen a la clase A y 10 ejemplos pertenecen a la

clase B, si nuestro modelo de clasificación siempre dice que son de la clase A, su precisión, o porcentaje de acierto, sería 99,9%. Totalmente engañoso si estamos interesados en que acierte en los ejemplos de la clase B ya que nunca detectaremos ningún ejemplo de la clase que interesa que se clasifique correctamente.

El porcentaje de acierto (*Accuracy*), para evaluar modelos de clasificación estándar, sí que es una buena medida, pero por lo explicado en el ejemplo anterior para la evaluación de modelos de clasificación no balanceada se usan otro tipo de medidas.

Para evaluar la calidad de estos clasificadores se usan métricas balanceadas en la *matriz de confusión*. Esta matriz se muestra en la tabla siguiente y en ella se anota el número de ejemplo mal y bien clasificado de ambas clases.

	Predicción positiva	Predicción negativa
Clase positiva	Verdadero Positivo (TP)	Falso Negativo (FN)
Clase negativa	Falso Positivo (FP)	Verdadero Negativo (TN)

Tabla 2 Matriz de confusión

A partir de la matriz de confusión, se definen métricas para el rendimiento para este tipo de problemas como las siguientes:

$Accuracy = \frac{TP+TN}{TP+FN+FP+TN}$ Porcentaje de ejemplos clasificados correctamente

$TP_{rate} = \frac{TP}{TP+FN}$ Porcentaje de ejemplos positivos clasificados correctamente

$TN_{rate} = \frac{TN}{FP+TN}$ Porcentaje de ejemplos negativos clasificados correctamente

$FP_{rate} = \frac{FP}{FP+TN}$ Porcentaje de ejemplos negativos mal clasificados

$FN_{rate} = \frac{FN}{TP+FN}$ Porcentaje de ejemplos positivos mal clasificados

Utilizar estas medidas de rendimiento no es apropiado si estamos interesados en obtener una clasificación de calidad parara ambas clases. El *Accuracy* por su limitación explicada en la sección anterior y el resto porque únicamente tienen en cuenta una clase concreta. Existen otras medidas que consisten combinar éstas como Kappa y media geométrica (GM) que paso a definir:

$Kappa = \frac{Acc - Acc_{random}}{1 - Acc_{random}}$ Donde Acc_{random} es:

$$Acc_{random} = \frac{(TN + FP) * (TN + FN) + (FN + TP) * (FP + TP)}{N * N}$$

$$GM = \sqrt{TP_{rate} * TN_{rate}}$$

Además de estas medidas, otra forma de combinar estas a partir de una observación de la curva ROC. Esta curva se presenta en un gráfico que permite visualizar la relación entre elTP_{rate} (beneficio de acertar) en el eje Y, frente al FP_{rate} (coste de fallar ejemplos positivos) en el eje X.

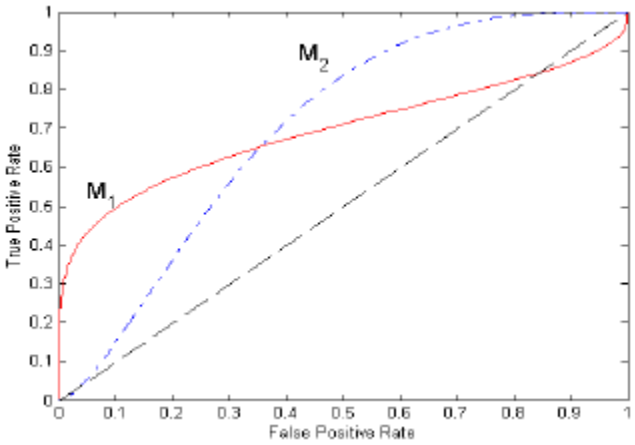

Gráfico 1 Ejemplo de dos curvas ROC de los clasificadores M1 y M2. La diagonal representa un clasificador aleatorio

El área bajo la curva ROC (AUC) nos provee de una medida de rendimiento para la clasificación del modelo que mejor sea en media. El gráfico 1 muestra dos curvas de dos modelos M1 y M2. En estas gráficas, el punto (0,0) y (1,1) son clasificaciones donde siempre se predice la clase negativa y positiva, respectivamente, y el punto (0,1) representa la clasificación perfecta. El área bajo la curva (AUC) se calcula de la siguiente forma:

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2}$$

En la literatura [5] podemos encontrar múltiples técnicas empleadas para hacer frente a problemas de clasificación (tanto clasificación estándar como no balanceada) como NaïveBayes, SVMs (Support Vector Machines), ó arboles de decisión entre otras. Entre las más utilizadas encontramos los Sistemas de Clasificación Basados en Reglas Difusas.

2.3 Sistemas de Clasificación Basados en Reglas Difusas

Existen numerosas técnicas para construir un clasificador (ya sea para clasificación estándar, o para datos no balanceados). Entre ellas, los Sistemas de Clasificación Basados en Reglas Difusas (SCBRDs) son una herramienta muy conocida y utilizada en el campo de reconocimiento de patrones y en problemas de clasificación. Los SCBRDs han sido empleados en múltiples problemas del mundo real, incluyendo la domótica, detección de intrusiones anómalas, procesamiento de imagen, y en medicina, entre otros.

La clasificación de los ejemplos en los SCBRDs se realiza en base a unas reglas fácilmente comprensibles por los seres humanos, las cuales son aprendidas a partir del conjunto de entrenamiento. La interpretabilidad del modelo obtenido es precisamente una de sus ventajas, para lo que hace uso de etiquetas lingüísticas (bajo, medio, alto, etc.) en sus reglas. Este modelo es capaz de representar la imprecisión del lenguaje humano (por ejemplo: "fiebre bastante alta") empleando la lógica difusa. Gracias a este tipo de lógica, el sistema podrá representar una regla del tipo "si el sueldo es alto y la situación laboral es estable entonces conceder préstamo", donde en este caso "sueldo" y "situación laboral" son variables difusas y "alto" y "estable" son etiquetas lingüísticas. Estas etiquetas se modelan con funciones de pertenencia como se ha explicado en la sección 2.1.

Los dos componentes principales de los SCBRDs son los siguientes:

- *Base de Conocimiento:* Está compuesta por la Base de Reglas (BR) y la Base de Datos (BD), donde se almacenan las reglas y las funciones de pertenencia difusas, respectivamente.
- *Método de Inferencia Difusa:* Este es el mecanismo que clásica los ejemplos empleando la información almacenada en la Base de Conocimiento.

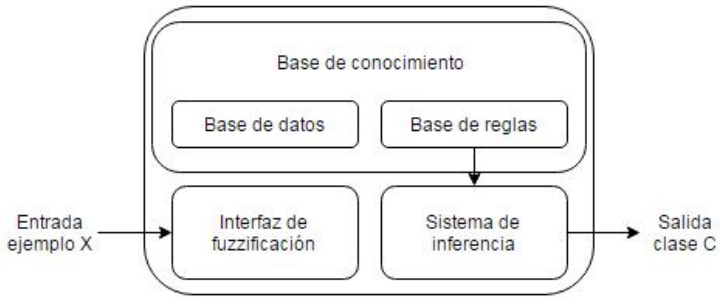

Ilustración 5 Estructura y componentes básicos de los SCBRDs

En la Ilustración 5 se muestran los principales componentes de los SCBRDs. Ante la entrada de un ejemplo, el sistema realiza un proceso de *fuzzificación* que consiste en transformar

los valores de las variables originales en valores difusos para poder aplicar sobre éste su base de conocimiento y sistema de inferencia para calcular la clase de salida.

En este proyecto empleamos uno de los SCBRDs más precisos e interpretables en la actualidad, FARC-HD y trataremos de mejorarlo usando técnicas conocidas y proponiendo nuevas técnicas. Además, realizaremos un estudio sobre el comportamiento de las técnicas de mejora en conjuntos de datos estándar y no balanceados.

2.3.1 FARC-HD

En este proyecto nos centramos en un algoritmo de aprendizaje de reglas difusas llamado FARC-HD *(Fuzzy Association Rule-based Classification model for High-Dimensional problems)*, presentado en [6] uno de los SCBRDs más precisos e interpretables de la literatura, además de muy estudiados [7]. Este algoritmo emplea la siguiente estructura de reglas:

Regla R_j: Si x_1 es A_{j1} y ... y x_n es A_{jn} entonces Clase = C_j *con* RW_j

Donde R_j es la etiqueta de la regla *j*-ésima, $x = (x_1; : : : ;x_n)$ es un vector *n*-dimensional que representa el ejemplo, A_{ji} es un conjunto difuso, $C_j \in C$ es la etiqueta de la clase *y* RW_j es el peso de la regla, el cual es calcula de la siguiente forma:

$$RW_j = CF_j = \frac{\sum_{x_p \in Clase C_j} \mu_{A_j}(x_p)}{\sum_{p=1}^{p} \mu_{A_j}(x_p)}$$

Donde $\mu_{Aj}(x_p)$ es el grado de emparejamiento del ejemplo x_p con los antecedentes de la regla difusa R_j , calculado mediante la ecuación (que se muestra más abajo). En el caso de FARC-HD, las etiquetas lingüísticas se modelan empleando funciones de pertenencia triangulares uniformemente distribuidas, las cuales forman una partición fuerte como se muestra en la Ilustración 6.

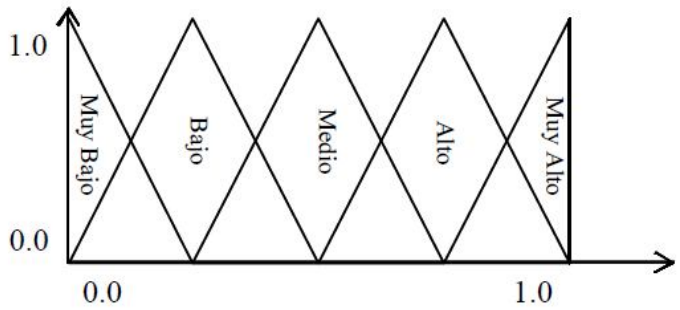

Ilustración 6 Ejemplo de las funciones de pertenencia de FARC-HD

A la hora de generar la Base de Reglas, FARC-HD aplica un proceso de aprendizaje compuesto de 3 etapas:

1. *Extracción de reglas de asociación difusas para clasificación*: Se emplea un árbol de búsqueda para cada clase mediante el que se obtienen las reglas difusas, cuyo número de antecedentes es limitado por la profundidad máxima del árbol.
2. *Filtrado de las reglas candidatas*: En esta fase se seleccionan las reglas más interesantes entre las obtenidas en la fase anterior por medio de un esquema de ejemplos ponderados.
3. *Selección evolutiva de reglas:* Se emplea un algoritmo evolutivo para ajustar los conjuntos difusos y seleccionar las reglas más precisas a partir de la base de reglas generada en la etapa anterior.

Una vez generadas las reglas difusas, si se debe clasificar un ejemplo $x_p = (x_{p1},..., x_{pn})$, FARC-HD aplica un Método de Razonamiento Difuso denominado combinación aditiva, compuesto por las siguientes fases:

1. *Grado de emparejamiento*. En esta fase se calcula la fuerza de activación de los antecedentes de todas las reglas de la BR con el patrón x_p, aplicando una *t-norma* T:

$$\mu_{A_j}(x_p) = T\left(\mu_{A_{j1}}(x_{p1}), \ldots, \mu_{A_j n_j}\left(x_{pn_j}\right)\right) \quad 1 \leq j \leq l$$

2. *Grado de asociación*. El grado de asociación del patrón x_p con cada regla de la BR se calcula como sigue:

$$b_j(x_p) = \mu_{A_j}(x_p) \cdot RW_j$$

3. *Grado de confianza.* En esta fase se calcula el grado de confianza para cada clase. Para esto, se suman los grados de asociación de las reglas de esa clase, es decir aquellas cuya clase consecuente es la clase que se está considerando.

$$conf_l(x_p) = \sum_{R_j \in RB;\, C_j = l} b_j(x_p) \quad , \quad l = 1,2,\ldots,m$$

Donde *m* es el número de clases.

4. *Clasificación*. Se predice aquella clase que obtenga el grado de confianza más alto.

$$Clase = arg \max_{l=1,\ldots,m} (conf_l(x_p))$$

2.4 Técnicas de creación de ensembles

El principal objetivo de estas técnicas consiste en tratar de mejorar el rendimiento de clasificadores individuales al introducir varios clasificadores y combinarlos para obtener un nuevo clasificador que supere a cada uno de ellos. Por tanto la idea básica es construir varios clasificadores a partir de los datos de entrenamiento originales y después, en la clasificación agregar las predicciones del conjunto de clasificadores base.

Esta idea sigue el comportamiento humano natural que tiende a tener en cuenta varias opiniones antes de tomar cualquier decisión importante. La principal motivación para la combinación de clasificadores en ensambles es mejorar su capacidad de generalización: cada clasificador se sabe que comete errores, cada uno diferentes, (ya que han sido entrenados en diferentes conjuntos de datos, se consideran clasificadores variantes al original) pero al usar el número adecuado de clasificadores base, estos errores pueden compensarse con aciertos y así mejorar la capacidad de generalización del clasificador original.

En la literatura, a pesar de no estar teóricamente claramente definido, la diversidad entre clasificadores es crucial (pero no es suficiente) para formar un buen ensemble. Tampoco se demuestra exactamente que la medida de de la diversidad esté relacionada con el *accuracy* (porcentaje de acierto del clasificador).

Hay diferentes formas de alcanzar la diversidad requerida, un punto importante es que cada uno de los clasificadores base deben ser *clasificadores débiles (weak learners)*, es decir que sus reglas se han aprendido con algoritmos de aprendizaje débil. Un algoritmo de aprendizaje se considera débil cuando los pequeños cambios en los datos de entrada puede produce grandes cambios en el modelo introducido.

Dos de las técnicas más usadas en el aprendizaje de modelos de ensemble son el Bagging y el Boosting. Ambas técnicas proporcionan una manera en la cual los clasificadores se generan estratégicamente para alcanzar la diversidad necesaria, mediante la manipulación del conjunto de entrenamiento para cada clasificador que formará el ensemble.

2.4.1 Bagging

Breiman [8] introdujo el concepto de *bootstrap* para la construcción de ensembles. Esto consiste básicamente entrenar los diferentes clasificadores base que formarán elensemble con replicas b*ootstrap* del conjunto de entrenamiento original. Así se crea un nuevo conjunto de entrenamiento para cada clasificador, el cual consiste en una muestra aleatoria (con reemplazo)

del conjunto de entrenamiento original. Además, generalmente se mantiene el tamaño del conjunto de entrenamiento original.

Por tanto, con cada réplica *Bootstrap* se aprende un clasificador base y debido a que el ensemble estará compuesto por tantos clasificadores base como réplicas se consideren. Cuando se desea clasificar un nuevo ejemplo, esté se clasifica con cada clasificador base individualmente. Se utiliza una mayoría de votos en las clases existentes, y cada clasificador puede votar a una o a varias clases, de esta forma diferenciaremos dos tipos de voto:

- *Voto Simple*: Cada clasificador base aporta un voto a la clase que predice, y la clase seleccionada es la clase que más votos obtenga.
- *Voto ponderado:* Cada clasificador base puede votar a todas las clases, en este caso el voto que cada clasificador aporta a cada clase corresponde al grado de confianza de cada clase explicado anteriormente en el método de razonamiento difuso explicado en la sección 2.3.1.

En el caso de empate de dos o más clases, se selecciona la clase mayoritaria entre las empatadas.

A continuación, se presenta un algoritmo de Bagging que realizaría un voto ponderado entre dos clases:

Entrada *S*: conjunto de entrenamiento; *T*: numero de iteraciones; *n*: Tamaño de la muestra bootstrap; *I*: clasificador base

Salida: Clasificador Bagging $H(x) = sign\left(\sum_{t=1}^{T} h_t(x)\right)$ donde $h_t \in [-1,1]$ corresponde a la hipótesis de cada clasificador base.

$For\ t = 1\ to\ T\ do$

$S_t \leftarrow MuestreoBootstrap(n, S)$

$h_t \leftarrow I(S_{t)}$

$End\ for$

2.4.2 Boosting

El concepto de Boosting fue introducido por Schapire en 1990. El algoritmo más representativo de esta familia es AdaBoost. Este algoritmo utiliza todo el conjunto de datos para aprender clasificadores en serie, pero después de cada ronda, da más enfoque a casos difíciles, con el objetivo de clasificar correctamente ejemplos en la siguiente iteración que fueron incorrectamente clasificados durante la iteración actual. Este algoritmo no lo estudiaremos

durante este proyecto, pero abre futuras líneas de estudio para realizar estudios similares sobre el Boosting.

A la hora de clasificar una nueva instancia se pueden usar igualmente los dos tipos de voto descritos en el apartado anterior.

3. Propuestas

En este proyecto, la principal propuesta es el desarrollo de un ensemble, usando la técnica de bagging, compuesto por SCBRDs como clasificadores base. El objetivo es estudiarla supuesta mejora que se obtiene en el rendimiento respecto al SCBRD que se usa para formar cada componente o clasificador base del ensemble. En nuestro caso emplearemos como SCBRD para formar el ensemble uno de los clasificadores más interpretables y precisos desarrollados en la actualidad: FARC-HD. Además, se propondrá una técnica para construir un nuevo clasificador a partir de toda la información que supone el aprendizaje de todas las bases de reglas y bases de datos que formarán el ensemble.

En los siguientes apartados, se describirá el modo en el que se ha construido el ensemble de FARC-HDs con la técnica bagging. Además, se describirá una técnica para obtener un nuevo clasificador a partir de todas las Bases de Datos y de Reglas que genera el aprendizaje de un ensemble.

3.1 Construcción de un ensemble de clasificadores FARC-HD con Bagging

Como hemos dicho en la sección 1.1, para construir un ensemble lo que se hace es entrenar un conjunto de clasificadores con varios conjuntos da datos extraídos a partir de un pre-procesamiento del conjunto de entrenamiento común la base de datos original. En concreto, nosotros vamos a utilizar Bagging y por tanto el número de clasificadores será un parámetro del modelo.

Para generar cada clasificador se debe seleccionar un subconjunto de ejemplos del conjunto de entrenamiento con el que realizar el aprendizaje del clasificador base. Para obtener este subconjunto se aplica *bootstrap aggregation*. Está basado en el muestreo aleatorio con reemplazamiento, para obtener muestras del mismo tamaño que el conjunto de datos original. Si nuestro conjunto de datos original tiene N elementos, realizaremos N extracciones para tener una muestra (un *bag*, saco). La probabilidad de que un ejemplo sea elegido para una muestra es *1/N* entonces la probabilidad de no ser elegido es *1- (1/N)*. Por tanto la probabilidad de que un ejemplo no sea escogido en N extracciones es $(1-(1/N))^N \approx e^{-1} = 0.368$. De esta forma cada muestra tendrá aproximadamente el 63.2% de los ejemplos del conjunto original (diferentes entre sí), y el resto serán ejemplos repetidos.

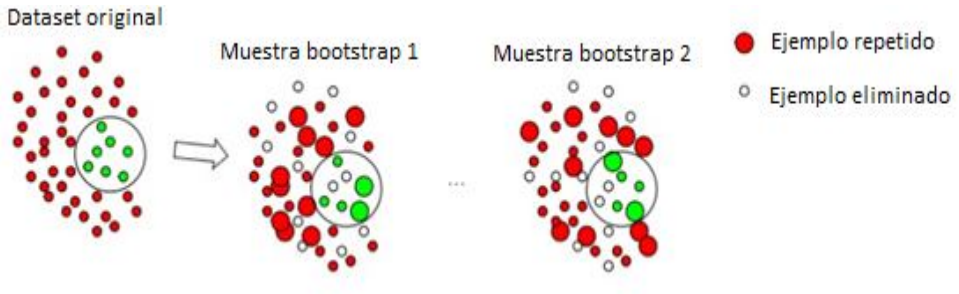

Ilustración 7 Muestreo bootstrap

En la ilustración 7 se muestra el funcionamiento del muestreo *bootstrap*. En este caso se observa que el dataset se compone de 2 clases (verde y roja) y una de ellas tiene un menor número de ejemplos. En la muestra 1 se observa que hay ciertos ejemplos que no se han seleccionado (circulo vacio) y otros se han seleccionado 2 o más veces (circulo ampliado), esto es debido al muestreo con reemplazamiento utilizado. Además cada una de las muestras es diferente.

Una vez obtenidas los *bags* se aplica FARC-HD original sobre cada una de las muestras (*bags*), de esta forma obtendremos tantos clasificadores como *bags* realicemos. Por lo tanto, tendremos tantas bases de datos y bases de reglas como *bags*. Además las reglas de cada base de reglas serán diferentes entre sí puesto que los ejemplos utilizados para aprenderlas son diferentes. El número de *bags* será un parámetro de estudio en este proyecto y se le denomina el tamaño del ensemble.

El objetivo de establecer el tamaño adecuado del ensemble, es obtener diversidad en los clasificadores base de esta forma se crearán grupos de ellos que serán más específicos para los diferentes tipos de ejemplos. Se propone realizar estudios con ensembles de tamaño 10 y 40.

A la hora de que el ensemble tenga que clasificar un nuevo ejemplo, se tienen en cuenta las predicciones de todos clasificadores para determinar la nueva clase. Como se ha explicado en la Sección 2.4.1 hay dos formas de hacerlo mediante una mayoría de votos:

- *Voto Simple*: Cada clasificador suma un voto a la clase predicha, y la clase seleccionada es la clase que más votos obtenga. En este caso la clase predicha por cada clasificador FARC-HD.
- *Voto ponderado:* Cada clasificador vota a todas las clases, pero a diferencia del voto simple, éste no suma uno a una clase, sino que cada clasificador aporta a cada clase la cantidad correspondiente al grado de confianza que se asocia a cada clase en el método de razonamiento difuso.

3.1.1 Implementación

Para la implementación de la propuesta, a partir del algoritmo original de FARC-HD descargado de Keel, se ha creado una nueva clase Java encargada de realizar el muestreo aleatorio, las llamadas correspondientes al FARC-HD que construirán el ensemble y la implementación de ambos tipos de voto. En la fase de clasificación FARC-HD ofrece una salida de una única clase con un grado de confianza. Para la implementación del voto simple, no supone problema ya que cada clase recibirá un voto por cada clasificador base. En cambio para la implementación del voto ponderado se ha modificado esta parte de la inferencia del algoritmo original para que cada clasificador base aporte a cada clase, la cantidad (a modo de voto) correspondiente al grado de confianza que se asocia a cada clase.

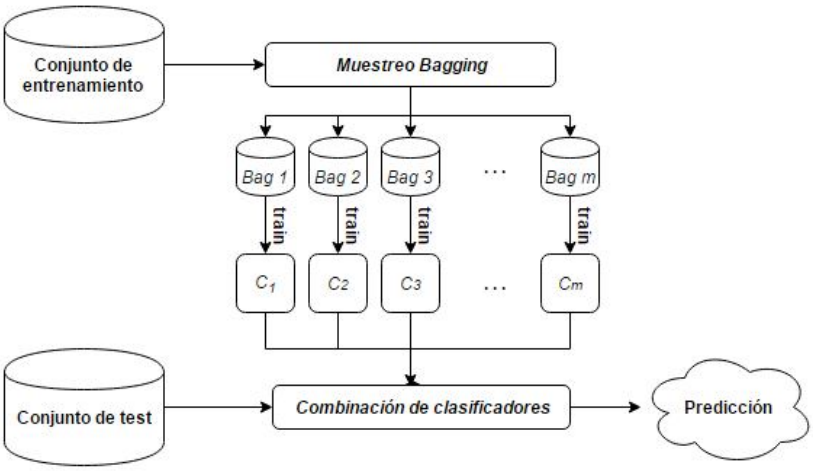

Ilustración 8 Funcionamiento ensemble con bagging

A modo de resumen, en la imagen anterior se muestra de forma general el funcionamiento de ensembles usando bagging.

Es necesario destacar que en la tercera fase del aprendizaje de FARC-HD original se hace uso de un algoritmo evolutivo para selección de reglas y ajuste de las funciones de pertenencia. Este algoritmo utiliza una heurística (*fitness*), para evaluar y optimizar los conjuntos de reglas seleccionadas. Para ello utiliza el *accuracy rate* pero para los problemas no balanceados se ha utilizado el AUC en un punto.

3.2 Creación de un clasificador difuso a partir de las bases de reglas de los clasificadores base del ensemble

La ejecución de bagging usando FARC-HD provoca la creación de diferentes Bases de Reglas y Bases de Datos, tantas como el tamaño del ensemble. Cada una de ellas, se ha aprendido a partir de una muestra bootstrap del conjunto de entrenamiento original. La última propuesta de este proyecto se basa en la extracción de reglas de todos estos clasificadores para formar un nuevo clasificador.

Por esto, y como parte de investigación se propondrán técnicas de selección de reglas y de las bases de datos asociadas a estas con el objetivo de hacer una fusión de todas bases de reglas y lograr construir un nuevo clasificador que también se comparará frente a los resultados que ofrece FARC-HD original. Esta parte es totalmente experimental y por eso proponemos un método simple a modo de abrir una nueva línea de investigación y se plantearán varias líneas futuras en base a la propuesta preliminar planteada.

3.2.1 Eliminación de reglas redundantes

La primera propuesta para procesar toda la información que tenemos tras la ejecución de Bagging está basada en la selección de reglas más generalistas ó la eliminación de reglas demasiado específicas (se puede interpretar de ambas formas). El proceso que deberíamos seguir para aplicar esta técnica consta de 3 fases:

1. Unimos el conjunto de todas las Bases de Reglas en una única. Esto provoca que en el nuevo conjunto de reglas existan muchas redundancias. Los siguientes apartados consisten en eliminar estas redundancias.
2. Eliminación de reglas repetidas: Puede que existan reglas con mismo antecedente y misma clase, entonces en el nuevo conjunto de reglas solo deberán aparecer una vez, sin repetidas. En el caso que existan reglas con el mismo antecedente pero diferente clase, seleccionaremos la regla con mayor grado de confianza.
3. Eliminación de reglas incluidas en otras: Para cada regla se comprueba si no está incluida en otra regla, si esto se cumple, entonces la regla correspondiente formará parte del nuevo clasificador. Es decir en nuestro nuevo clasificador deberemos incluir únicamente las reglas que no estén incluidas en otras.

Para aclarar la fase 3 de eliminación las reglas que incluyen a otras, se va a mostrar un ejemplo. Una vez eliminadas las reglas repetidas de la unión de todas las bases de reglas se

obtiene el siguiente conjunto de reglas, al que le aplicaremos la fase de eliminación de reglas incluidas en otras.

```
1: petalWidth  IS L_0(5): Iris-setosa
2: petalLength IS L_0(5): Iris-setosa
3: petalLength IS L_2(5): Iris-versicolor
4: petalWidth  IS L_4(5): Iris-virginica
5: petalWidth  IS L_3(5): Iris-virginica
6: petalLength IS L_3(5): Iris-virginica
7: petalLength IS L_4(5): Iris-virginica
8: petalWidth  IS L_2(5): Iris-versicolor
9: petalLength IS L_3(5) AND petalWidth IS L_3(5): Iris-virginica
10: sepalWidth IS L_1(5) AND petalWidth IS L_3(5): Iris-virginica
```

Para cada par de reglas se comprueba si una está incluida en la otra y una regla se mantendrá en nuestra nueva base de reglas si no está incluida por ninguna otra. Si nos fijamos en las reglas, las que tienen el antecedente de la regla formado por una única composición IS es imposible que estén incluidas en otras (serian repetidas y en este ejemplo ya se han eliminado las repetidas). Por tanto nos fijamos en las reglas 9 y 10.

La regla 9 está incluida tanto en la regla 5 como en la 6 ya que los antecedentes de ambas aparecen en la regla 9. Si nos fijamos en la regla 10, también está cubierta (incluida) por la regla 6. Por consiguiente, en este ejemplo se seleccionan para el nuevo clasificador las reglas 5 y 6 ya que son más generales que la 9 y la 10 las cuales se descartan. Entonces la base de reglas final sería la formada por la1 hasta la regla 8 incluida.

En este punto, tenemos formada la Base de Reglas de nuestro nuevo clasificador, pero tenemos un problema. Cada una de estas reglas seleccionadas se ha calculado con diferentes Bases de Datos (aunque muy similares). Por lo tanto también debemos seleccionar la base de datos que usaremos para nuestro nuevo clasificador.

Para ello se testea el rendimiento sobre el conjunto de datos original (sin muestrear) de cada FARC-HD aprendido en el ensemble, y una vez conocido el rendimiento que obtiene cada uno, seleccionamos para nuestro nuevo clasificador la base de datos del FARC-HD que mejor rendimiento obtenga sobre el conjunto de datos original. Al realizar esta evaluación de los clasificadores base sobre el conjunto de entrenamiento original (sin muestrear), como hemos dicho en la Sección 3.1, teóricamente el 63.2% de ellos se ha seleccionado para el aprendizaje del clasificador base y el otro 36.8% son aquellos ejemplos que no se seleccionaron para la muestra del aprendizaje. Con esto se pretende obtener una estimación más realista de cada uno de los clasificadores base.

4. Marco experimental

En este capítulo presentamos la configuración del marco experimental utilizado para desarrollar los experimentos llevados a cabo en la Sección 5. Primero, describimos los conjuntos de datos (datasets) seleccionados para el estudio experimental (Sección 4.1 y 4.2). Posteriormente, mostramos las medidas de rendimiento seleccionadas para cada conjunto de datos e introduciremos las pruebas estadísticas necesarias para comprobar si existen diferencias entre los resultados obtenidos (Sección 4.3)

4.1 Conjuntos de datos de clasificación estándar

Con el objetivo de analizar el rendimiento de nuestra propuesta, hemos considerado 18 conjuntos de datos (datasets) del mundo real seleccionados del repositorio de KEEL, los cuales están disponibles públicamente desde el sitio web del proyecto, la tabla siguiente resume las características de los conjuntos de datos seleccionados, indicando por cada uno el número de ejemplos (#Ej.), número de atributos totales (#Atr), además se especifica número de atributos reales (#AtrR), el número de atributos enteros (#AtrI), el número de atributos nominales (#AtrN), y número de clases (#Clas).

Para llevar a cabo los diferentes experimentos hemos considerado un modelo de validación cruzada de 5 particiones, es decir, dividimos el conjunto de datos en 5 particiones, cada una conteniendo un 20% de los ejemplos, y empleamos una combinación de cuatro de ellos (80%) para entrenar el sistema y el resto para probarlo. De esta manera, el resultado de cada conjunto de datos se obtiene calculando la media aritmética de las cinco particiones. En lugar de emplear la validación cruzada que se utiliza habitualmente y con el objetivo de corregir la fractura de los datos (*dataset shift*), es decir, cuando el conjunto de datos y el conjunto de prueba no siguen la misma distribución haremos uso de un procedimiento de particionado llamado *Distribution Optimally Balanced Standard Cross Validation* (DOB-SCV).

Nombre	#Ej	#Atr	#AtrR	#AtrI	#AtrN	#Clas
automobile	150	25	15	0	10	6
balance	625	25	15	0	10	3
cleveland	297	13	13	0	0	5
contraceptive	1473	9	0	9	0	3
ecoli	336	7	7	0	0	8
glass	214	9	9	0	0	7
hayes-roth	160	4	0	4	0	3
iris	150	4	4	0	0	3
newthyroid	215	5	4	1	0	3
page-blocks	5472	10	4	6	0	5
segment	2310	19	19	0	0	7
shuttle	58000	9	0	9	0	7
tae	151	5	0	5	0	3
thyroyd	7200	21	6	15	0	3
vehicle	846	18	0	18	0	4
vowel	990	13	10	3	0	11
wine	178	13	13	0	0	3
yeast	1484	8	8	0	0	10

Tabla 3 Características de los conjuntos de datos estándar considerados en el estudio experimental

4.2 Conjuntos de datos no balanceados

De la misma forma que los datasets estándar, obtenemos 22 conjuntos de datos no balanceados del sitio web del proyecto KEEL. A continuación, se expondrá una tabla resumen para describir los conjuntos de datos de forma similar a que hemos descrito los datasets anteriores.

En el caso de los no balanceados, como únicamente existen dos clases, es necesario indicar que el ratio de no balanceo *(Imbalanced Rate IR)* en lugar del número de clases. Este valor se calcula dividiendo el numero de ejemplos de la clase mayoritaria entre el de la clase minoritaria y cuanto mayor es, más desbalanceado esta. En este conjunto de datasets varía entre 1.5 y 9. Además se han obtenido los datasets ya particionados en 5 particiones mediante *Folder Cross Validation (FCV)*.

Nombre	#IR	#Ej	#Atr	#AtrR	#AtrI	#AtrN
ecoli-0_vs_1-5	1.86	220	7	7	0	0
ecoli1-5	3.36	336	7	7	0	0
ecoli2-5	5.46	336	7	7	0	0
ecoli3-5	8.6	336	7	7	0	0
glass-0-1-2-3_vs_4-5-6-5	3.2	214	9	9	0	0
glass0-5	2.06	214	9	9	0	0
glass1-5	1.82	214	9	9	0	0
glass6-5	6.38	214	9	9	0	0
haberman-5	2.78	306	3	0	3	0
iris0-5	2	150	4	4	0	0
new-thyroid1-5	5.14	215	5	4	1	0
new-thyroid2-5	5.14	215	5	4	1	0
page-blocks0-5	8.79	5472	10	8	6	0
pima-5	1.87	768	8	19	0	0
segment0-5	6.02	2308	19	0	0	0
vehicle0-5	3.25	846	18	0	18	0
vehicle1-5	2.9	846	18	0	18	0
vehicle2-5	2.88	846	18	0	18	0
vehicle3-5	2.99	846	18	0	18	0
wisconsin-5	1.86	683	9	0	9	0
yeast1-5	2.46	1484	8	8	0	0
yeast3-5	8.1	1484	8	8	0	0

Tabla 4 Características de los conjuntos de datos no balanceados considerados en el estudio experimental

4.3 Medidas de rendimiento y comparación de clasificadores

En este proyecto, aparte de la implementación del Bagging, gran trabajo se centra en el estudio de la calidad de los clasificadores, para esto es necesario realizar una comparación honesta de la bondad del clasificador. Para evaluar honestamente un clasificador es necesario definir una medida que nos indique la calidad del clasificador.

Como se ha explicado en las secciones 2.1 y 2.2 existen diferentes tipos de medidas de rendimiento para calcular la precisión o exactitud, en este proyecto haremos un estudio sobre el comportamiento del ensemble creado con Bagging frente al FARC-HD original en diferentes tipos de datos. En concreto para la clasificación estándar, usaremos para las comparaciones, como medida de rendimiento el *accuracy* (tasa de acierto). Pero para la clasificación de datos no balanceados usaremos diferentes medidas debido a la limitación explicada. Usaremos la media geométrica, el Kappa y el AUC.

En este estudio, para las comparaciones de clasificadores realizaremos un análisis estadístico para realizar una comparativa honesta entre los clasificadores. Usaremos el test de Wilcoxon, en apartado de estudio experimental para realizar comparativas entre las diferentes configuraciones del Bagging variando el número de clasificadores base empleados (el tamaño del ensemble) y las dos técnicas para combinar las predicciones de todos los clasificadores base para obtener la predicción del ensemble (el tipo de voto).

El test de rangos de Wilcoxon es un test estadístico que permite la realizar una comparación objetiva de los resultados de dos clasificadores en varios datasets. La información que aporta es si existen diferencias estadísticas entre un par de clasificadores.

Otro tipo de test para la comparación de clasificadores es la prueba de rangos alineados de Friedman. A diferencia de Wilcoxon, éste test permite la comparación de varios clasificadores y varios datasets (Wilcoxon permite comparar únicamente los resultados de dos clasificadores). El test de rangos alineados de Friedman asocia a cada clasificador un valor (su rango) y el mejor clasificador es el de menor rango, al contrario que Wilcoxon que asocia el mayor rango al mejor. El clasificador que obtiene el menor rango es el mejor y se le denomina método de control. Además esta prueba calcula un p-valor que nos indicará si existen diferencias significativas entre el clasificador del menor rango y el resto.

5. Estudio experimental: Bagging

5.1 Estudio sobre problemas de clasificación estándar

En esta sección se presenta un análisis de los resultados de los clasificadores generados con el Bagging. Primero se realizara una comparativa de los resultados de los clasificadores del mismo tamaño con el test de Wilcoxon para observar si existen diferencias usando ambos tipos de votos, y a continuación, si ha habido diferencias entre los clasificadores del mismo tamaño, nos quedaremos con el que mejores resultados ofrezca, si no hay diferencias, seleccionaremos el que mayor rango ofrezca en el test.

Dataset	Bagging			
	Voto simple		Voto ponderado	
	10	40	10	40
automobile	74.14	76.88	72.55	76.29
balance	89.45	90.88	88.97	91.2
cleveland	57.52	59.9	59.9	59.89
contraceptive	54.72	54.24	54.92	54.58
ecoli	85.79	86.64	86.67	86.08
glass	64.45	67.79	64.91	66.78
hayes-roth	80.03	80.07	84.41	82.49
Iris	95.33	95.33	95.33	95.33
newthyroid	96.74	96.74	97.21	96.74
page-blocks	95.16	95.05	95.29	95.21
segment	93.12	94.07	93.38	93.81
shuttle	96.94	96.86	96.31	96.31
tae	59.82	57.73	56.29	57.04
thyroid	94.39	94.44	94.38	94.35
vehicle	69.88	70.81	72.46	72.11
vowel	79.49	81.82	79.09	80.4
wine	95.57	97.78	97.78	96.67
yeast	59.3	60.45	59.58	59.57
Media	80.1	80.97	80.52	80.82

Tabla 5 Resultados de ensembles en clasificación estándar

En la Tabla 5, en cada fila se muestra el porcentaje de aciertos en test (*accuracy*) de los cuatro ensembles implementados sobre diferentes conjuntos de datos. Cada dato es el promedio de las 5 particiones en las que se divide cada dataset. Para cada dataset se muestra en verde la mejor tasa de acierto y en rojo la peor y en la última fila se muestra la media de cada columna.

Por un lado, si nos fijamos en las columnas de 10FARC-HD se observa que con el voto simple únicamente obtiene los mejores resultados en dos datasets, en cambio usando el voto ponderado obtiene el mejor resultado en 8 conjuntos de datos, lo que lleva a pensar que puede ser una de las mejores configuraciones.

Por otro lado, las columnas de 40FARC-HD se observa que el voto simple obtiene muchos mejores resultados ya que tiene el mejor dato en 9 datasets en cambio el voto ponderado únicamente lo tiene en uno. En la Tabla 6 se muestra un resumen del número de datasets sonde se obtiene el mejor y peor resultado en todas las configuraciones. Además, en media el ensemble que mejor resultado ofrece es con 40 FARC-HDs combinando sus resultados con el voto simple.

Configuración	Datasets con mejor resultado	Datasets con peor resultado
Tamaño 10 voto simple	2	10
Tamaño 10 voto ponderado	8	4
Tamaño 40 voto simple	9	4
Tamaño 40 voto ponderado	1	4

Tabla 6 Resumen de resultados con el *accuracy*

En las siguientes secciones se realizará un análisis estadístico para lograr una comparación más honesta y objetiva de los resultados anteriores.

5.1.1 Influencia del tipo de voto

En esta sección se va a determinar para cada tamaño que tipo de voto ofrece mejores resultados. Observando los resultados medios, parece que para un ensemble de tamaño 10 ofrece mejores resultados con el voto ponderado, pero únicamente un 0.42% mejor. En cambio con tamaño 40 ofrece mejor promedio el voto simple pero en este caso también solo un 0.15% mejor. Estas conclusiones no son muy objetivas ya que al existir tan poca diferencia entre las medias, añadir nuevos conjuntos de datos al estudio esto podría variar.

A continuación, en la Tabla 7, se muestran los resultados del test de Wilcoxon comparando ambos tipos de votos para los dos tamaños del ensemble.

Comparación	R+	R-	p-valor	Hipótesis α=0.05	Selección
S10FARC vs P10FARC	60.5	110.5	0.2868	Aceptado	**P10FARC**
S40FARC vs P40FARC	115.5	55.5	0.2146	Aceptado	**S40FARC**

Tabla 7 Wilcoxon para el tipo de voto usando el *accuracy*

Debido a que ningún p-valor es menor que 0.05 no existen diferencias significativas entre cada par de clasificadores, pero, ya que los mejores resultados en medias coinciden con los mayores rangos, los seleccionados han sido P10FARC (10 clasificadores con voto ponderado) y S40FARC (40 clasificadores con voto simple).

5.1.2 Influencia del tamaño del ensemble

En base al estudio anterior se va a determinar cuál es la mejor configuración que se usará para la comparación con FARC-HD original

Comparación	R+	R-	p-valor	Hipótesis α=0.05	Selección
P10FARC vs S40FARC	60	111	0. 2524	Aceptado	**S40FARC**

Tabla 8 Wilcoxon para el tamaño según la Tabla 7

El resultado del test nuevo nos ofrece que no existen diferencias entre ambos clasificadores, pero al obtener el mayor rango el de 40 clasificadores con voto simple y ofrecer los mejores resultados en cuanto al número de datasets que mejor resultado ofrece, es el que usaremos en la comparación con FARC-HD original. Si tuviésemos que decidirnos por uno sería el de 40 clasificadores con voto simple, pero no se ha podido comprobar que existan diferencias significativas con el resto.

5.1.3 Comparación con FARC-HD original

En la Tabla 9 se ha añadido los resultados del algoritmo original (FARC) con los resultados de los ensembles mostrados en la Tabla 5.

Dataset	FARC	Bagging			
		Voto simple		Voto ponderado	
		10	40	10	40
automobile	73.68	74.14	76.88	72.55	76.29
balance	87.53	89.45	90.88	88.97	91.20
cleveland	58.89	57.52	59.90	59.90	59.89
contraceptive	53.97	54.72	54.24	54.92	54.58
ecoli	84.60	85.79	86.64	86.67	86.08
glass	65.02	64.45	67.79	64.91	66.78
hayes-roth	78.80	80.03	80.07	84.41	82.49
iris	94.67	95.33	95.33	95.33	95.33
newthyroid	93.95	96.74	96.74	97.21	96.74
page-blocks	95.27	95.16	95.05	95.29	95.21
segment	92.90	93.12	94.07	93.38	93.81
Shuttle	97.48	96.94	96.86	96.31	96.31
tae	57.84	59.82	57.73	56.29	57.04
thyroid	94.26	94.39	94.44	94.38	94.35
vehicle	68.45	69.88	70.81	72.46	72.11
vowel	74.95	79.49	81.82	79.09	80.40
wine	95.54	95.57	97.78	97.78	96.67
yeast	59.65	59.30	60.45	59.58	59.57
Media	79.30	80.10	80.97	80.52	80.82

Tabla 9 Resultados de FARC-HD original y resultados de ensembles en clasificación estándar

A simple vista se observa en la tabla anterior que FARC-HD original tiene una media peor que los ensembles ya que prácticamente en todos los datasets obtiene el peor resultado.

En las secciones anteriores, los test estadísticos no nos han aportado información suficiente como para poder comparar únicamente el mejor ensemble con FARC original. Por esto se ha decidido hacer la prueba de Wilcoxon con todas las configuraciones, los resultados se muestran en la siguiente tabla.

Comparación	**R+**	**R-**	**p-valor**	**Hipótesis α=0.05**	**Selección**
FARC vs S10FARC	35	136	0.0279	Rechazado	**S10FARC**
FARC vs P10FARC	37	134	0.0347	Rechazado	**P10FARC**
FARC vs S40FARC	9	162	0.0009	Rechazado	**S40FARC**
FARC vs P40FARC	19	152	0.0038	Rechazado	**P40FARC**

Tabla 10 Wilcoxon de todos ensembles con FARC-HD original

Según los estudios de las secciones 5.1.1 y 5.1.2 no se ha podido determinar que ningún ensemblesea diferente al resto, en este caso se comprueba estadísticamente que todas las configuraciones del bagging han superado a los resultados de FARC-HD original. En concreto, también se observa que la configuración con voto simple de 40 clasificadores es la que mayor rango ofrece de todos, por lo tanto la mejor.

5.2 Estudio sobre datos no balanceados

En las siguientes secciones se realizará el mismo estudio que se ha realizado en la sección anterior pero sobre conjuntos de datos no balanceados y con diferentes medidas de rendimiento (AUC, Kappa y Media geométrica) debido a la limitación del accuracy explicada en la Sección 2.2 para evaluar el rendimiento en este tipo de problemas.

5.2.1 AUC

En la Tabla 11 se muestran los resultados sobre datos no balanceados medidos con el AUC, siguiendo el formato de verde el mejor y rojo el peor por cada fila. Además en este estudio los resultados se muestran ordenados por IR (*Imbalanced Rate*) o ratio de no balanceo.

Dataset	IR	Bagging			
		Voto simple		Voto ponderado	
		10	40	10	40
glass1-5	**1.82**	0.735	0.746	0.737	0.771
ecoli-0_vs_1-5	**1.86**	0.983	0.983	0.983	0.98
wisconsin-5	**1.86**	0.963	0.97	0.968	0.971
pima-5	**1.87**	0.755	0.762	0.74	0.763
iris0-5	**2**	1	1	1	1
glass0-5	**2.06**	0.819	0.834	0.824	0.834
yeast1-5	**2.46**	0.714	0.729	0.718	0.723
haberman-5	**2.78**	0.618	0.632	0.636	0.632
vehicle2-5	**2.88**	0.949	0.954	0.955	0.955
vehicle1-5	**2.9**	0.701	0.702	0.707	0.705
vehicle3-5	**2.99**	0.704	0.704	0.701	0.716
glass-0-1-2-3_vs_4-5-6-5	**3.2**	0.876	0.926	0.899	0.912
vehicle0-5	**3.25**	0.94	0.951	0.95	0.955
ecoli1-5	**3.36**	0.882	0.878	0.865	0.876
new-thyroid1-5	**5.14**	0.957	0.969	0.952	0.98
new-thyroid2-5	**5.14**	0.971	0.969	0.954	0.969
ecoli2-5	**5.46**	0.896	0.903	0.907	0.901
segment0-5	**6.02**	0.988	0.994	0.989	0.984
glass6-5	**6.38**	0.878	0.878	0.878	0.881
yeast3-5	**8.1**	0.92	0.928	0.921	0.925
ecoli3-5	**8.6**	0.837	0.851	0.872	0.881
page-blocks0-5	**8.79**	0.878	0.878	0.877	0.873
Media		0.862	0.87	0.865	0.872

Tabla 11 Resultados de ensembles en clasificación no balanceada con AUC

Ante estos resultados, *a priori* parece que el ensemble que mejores resultados ofrece es el de tamaño 40 con voto ponderado pero esto lo comprobaremos en el siguiente estudio. A diferencia del estudio en la clasificación estándar, se observa que el voto ponderado en este

caso obtiene mejor media que los simples para ambos tipos de voto. En la Tabla 12 se muestra un resumen con el número de datasets que obtienen el mejor y el peor valor.

Configuración	Datasets con mejor resultado	Datasets con peor resultado
Tamaño 10 voto simple	4	15
Tamaño 10 voto ponderado	3	7
Tamaño 40 voto simple	4	2
Tamaño 40 voto ponderado	11	4

Tabla 12 Resumen de resultados con el AUC

En esta ocasión, el que mejores resultados ofrece es el de tamaño 40 con voto ponderado (P40FARC) a diferencia de en clasificación estándar en la que el que mejor resultado ofrece es el de 40 pero con voto simple.

5.2.1.1 Influencia con el tipo de voto

En la siguiente tabla se muestran los resultados del test de Wilcoxon de la comparación del tipo de voto en ensembles del mismo tamaño.

Comparación	R+	R-	p-valor	Hipótesis α=0.05	Selección
S10FARC vs P10FARC	85.5	167.5	0.1913	Aceptado	**P10FARC**
S40FARC vs P40FARC	108.5	144.5	0.6012	Aceptado	**P40FARC**

Tabla 13 Wilcoxon para el tipo de voto usando el AUC

Según los resultados de la Tabla 13, ya que ningún p-valor es inferior a 0.05 no existen diferencias entre los clasificadores del mismo tamaño (variando el voto). Por esta razón en el estudio de la influencia del tamaño, lo realizamos con los clasificadores que mayor rango han obtenido, en esta ocasión los que usan el voto ponderado.

5.2.1.2 Influencia del tamaño del ensemble

En el apartado anterior se ha comprobado que no había diferencias estadísticas entre los clasificadores del mismo tamaño (pero variando el voto). Por esta razón en este estudio se comprobará si existen diferencias entre los ensembles que usan el mismo voto (en este caso el ponderado ya que obtiene el mayor rango en las pruebas anteriores) pero variando el tamaño. Los resultados que obtenemos al comparar ambos ensembles con el voto ponderado se muestran en la siguiente tabla.

Comparación	R+	R-	p-valor	Hipótesis α=0.05	Selección
P10FARC vs P40FARC	48.5	204.5	0.0106	Rechazado	**P40FARC**

Tabla 14 Wilcoxon para el tamaño según la Tabla 13

En este caso ya que el p-valor es menor que 0.05, existen diferencias estadísticas entre ambos ensembles siendo mejor el de tamaño 40. Por esto en la siguiente fase compararemos únicamente este ensemble contra el FARC-HD original en lugar de compararlos todos.

5.2.1.3 Comparación con FARC-HD original

En la Tabla 15 se ha añadido los resultados del algoritmo original con los resultados de los ensembles mostrados en la Tabla 11.

Dataset	IR	FARC	Bagging			
			Voto simple		Voto ponderado	
			10	40	10	40
glass1-5	**1.82**	0.739	0.735	0.746	0.737	0.771
ecoli-0_vs_1-5	**1.86**	0.98	0.983	0.983	0.983	0.98
wisconsin-5	**1.86**	0.969	0.963	0.97	0.968	0.971
pima-5	**1.87**	0.753	0.755	0.762	0.74	0.763
iris0-5	**2**	1	1	1	1	1
glass0-5	**2.06**	0.792	0.819	0.834	0.824	0.834
yeast1-5	**2.46**	0.72	0.714	0.729	0.718	0.723
haberman-5	**2.78**	0.604	0.618	0.632	0.636	0.632
vehicle2-5	**2.88**	0.932	0.949	0.954	0.955	0.955
vehicle1-5	**2.9**	0.694	0.701	0.702	0.707	0.705
vehicle3-5	**2.99**	0.709	0.704	0.704	0.701	0.716
glass-0-1-2-3_vs_4-5-6-5	**3.2**	0.87	0.876	0.926	0.899	0.912
vehicle0-5	**3.25**	0.933	0.94	0.951	0.95	0.955
ecoli1-5	**3.36**	0.876	0.882	0.878	0.865	0.876
new-thyroid1-5	**5.14**	0.966	0.957	0.969	0.952	0.98
new-thyroid2-5	**5.14**	0.954	0.971	0.969	0.954	0.969
ecoli2-5	**5.46**	0.884	0.896	0.903	0.907	0.901
segment0-5	**6.02**	0.983	0.988	0.994	0.989	0.984
glass6-5	**6.38**	0.914	0.878	0.878	0.878	0.881
yeast3-5	**8.1**	0.921	0.92	0.928	0.921	0.925
ecoli3-5	**8.6**	0.819	0.837	0.851	0.872	0.881
page-blocks0-5	**8.79**	0.871	0.878	0.878	0.877	0.873
Media		0.858	0.862	0.87	0.865	0.872

Tabla 15 Resultados de FARC-HD original y resultados de ensembles en clasificación no balanceada con AUC

Según los análisis de las secciones 5.2.1.1 y 5.2.1.2 se ha determinado que la mejor configuración es con tamaño 40 y voto ponderado, por lo tanto a continuación se muestra el test de Wilcoxon comparando FARC-HD original con este. También podemos observar en la tabla que todos los ensembles superan en media al algoritmo original.

Comparación	R+	R-	p-valor	Hipótesis α=0.05	Selección
FARC vs P40FARC	24.5	228.5	0.001	Rechazado	**P40FARC**

Tabla 16 Wilcoxon mejor ensemble con FARC-HD original

Se comprueba que el bagging con 40 clasificadores y voto ponderado mejora estadísticamente al FARC-HD original usando el AUC como medida de rendimiento para el

clasificador. A continuación se seguirá utilizando la misma metodología para el resto de las medidas de rendimiento.

5.2.2 Kappa

Del mismo modo que se ha realizado el estudio con el AUC mostramos a continuación los resultados de los ensembles con Kappa para obtener la mejor configuración del ensemble.

Dataset	IR	Bagging			
		Voto simple		Voto ponderado	
		10	40	10	40
glass1-5	**1.82**	0.474	0.495	0.48	0.535
ecoli-0_vs_1-5	**1.86**	0.969	0.97	0.97	0.96
wisconsin-5	**1.86**	0.923	0.936	0.932	0.936
pima-5	**1.87**	0.494	0.507	0.467	0.51
iris0-5	**2**	1	1	1	1
glass0-5	**2.06**	0.646	0.657	0.611	0.65
yeast1-5	**2.46**	0.402	0.427	0.399	0.413
haberman-5	**2.78**	0.239	0.262	0.276	0.265
vehicle2-5	**2.88**	0.901	0.913	0.905	0.913
vehicle1-5	**2.9**	0.402	0.392	0.406	0.398
vehicle3-5	**2.99**	0.381	0.378	0.373	0.393
glass-0-1-2-3_vs_4-5-6-5	**3.2**	0.785	0.857	0.826	0.832
vehicle0-5	**3.25**	0.848	0.87	0.859	0.874
ecoli1-5	**3.36**	0.735	0.719	0.701	0.715
new-thyroid1-5	**5.14**	0.946	0.948	0.911	0.95
new-thyroid2-5	**5.14**	0.964	0.945	0.927	0.945
ecoli2-5	**5.46**	0.778	0.816	0.8	0.804
segment0-5	**6.02**	0.972	0.982	0.977	0.97
glass6-5	**6.38**	0.809	0.809	0.809	0.826
yeast3-5	**8.1**	0.744	0.734	0.72	0.733
ecoli3-5	**8.6**	0.566	0.582	0.653	0.635
page-blocks0-5	**8.79**	0.593	0.591	0.612	0.617
Media		0.708	0.718	0.71	0.721

Tabla 17 Resultados de ensembles en clasificación no balanceada con Kappa

De nuevo mostraremos una tabla resumen para analizaren cuantos conjuntos de datos obtiene cada ensemble el mejor y peor resultado.

Configuración	Datasets con mejor resultado	Datasets con peor resultado
Tamaño 10 voto simple	3	11
Tamaño 10 voto ponderado	4	10
Tamaño 40 voto simple	7	4
Tamaño 40 voto ponderado	9	3

Tabla 18 Resumen de resultados con el Kappa

En esta ocasión se observa que el voto ponderado está mejorando los resultados de los ensembles de ambos tamaños, es decir, pasar de tamaño 10 a 40 y de voto simple a ponderado en ambos casos se aumentan el número de datasets donde se obtiene el mejor resultado y hace que disminuya el número de peores resultados. Las mejoras son más significativas aumentando a 40 clasificadores, que pasando al voto ponderado.

Por esto, a la vista de los resultados de las tablas 17 y 18 se sigue manteniendo que los mejores resultados *a priori* se observan con el ensemble de tamaño 40 y voto ponderado, pero con el objetivo de comprobar estos resultados se ha seguido haciendo los correspondientes estadísticos mostrados a continuación.

5.2.2.1 Influencia con el tipo de voto

Del mismo modo que la Sección 5.2.1.1 se muestran los resultados de los test comparando los ensembles del mismo tamaño con los diferentes tipos de voto.

Comparación	R+	R-	p-valor	Hipótesis α=0.05	Selección
S10FARC vs P10FARC	114.5	138.5	0.7650	Aceptado	**P10FARC**
S40FARC vs P40FARC	109.5	143.5	0.5755	Aceptado	**P40FARC**

Tabla 19 Wilcoxon para el tipo de voto usando el Kappa

De nuevo, al igual que ha sucedido en la Sección 5.2.1.1 con el AUC, se comprueba que con el mismo tamaño no existen diferencias estadísticas entre los ensembles, por esto, y por que en ambos test el voto ponderado obtiene el mayor rango, para la siguiente sección se hará la comparativa con los ensembles que utilizan el voto ponderado.

5.2.2.2 Influencia del tamaño del ensemble

En la Tabla 20 se muestra la comparación de los ensembles que mejor rango han obtenido en el apartado anterior, al igual que con el AUC, el voto ponderado para ambos tamaños.

Comparación	R+	R-	p-valor	Z	Hipótesis α=0.05	Selección
P10FARCvsP40FARC	50.5	202.5	0.0143	-2.4504	Rechazado	**P40FARC**

Tabla 20 Wilcoxon para el tamaño según la Tabla 19

Se vuelve a comprobar mediante el Kappa que la mejor configuración del ensemble es con 40 clasificadores y usando el voto ponderado. Por esto en la siguiente sección se comparará únicamente los resultados de esta configuración con FARC-HD original.

5.2.2.3 Comparación con FARC-HD original

A continuación, se muestran los resultados del algoritmo original evaluado con el Kappa junto a los resultados de los ensembles de la Tabla 17.

Dataset	IR	FARC	Bagging			
			Voto simple		Voto ponderado	
			10	40	10	40
glass1-5	**1.82**	0.459	0.474	0.495	0.48	0.535
ecoli-0_vs_1-5	**1.86**	0.96	0.969	0.97	0.97	0.96
wisconsin-5	**1.86**	0.93	0.923	0.936	0.932	0.936
pima-5	**1.87**	0.485	0.494	0.507	0.467	0.51
iris0-5	**2**	1	1	1	1	1
glass0-5	**2.06**	0.573	0.646	0.657	0.611	0.65
yeast1-5	**2.46**	0.401	0.402	0.427	0.399	0.413
haberman-5	**2.78**	0.2	0.239	0.262	0.276	0.265
vehicle2-5	**2.88**	0.828	0.901	0.913	0.905	0.913
vehicle1-5	**2.9**	0.358	0.402	0.392	0.406	0.398
vehicle3-5	**2.99**	0.365	0.381	0.378	0.373	0.393
glass-0-1-2-3_vs_4-5-6-5	**3.2**	0.758	0.785	0.857	0.826	0.832
vehicle0-5	**3.25**	0.819	0.848	0.87	0.859	0.874
ecoli1-5	**3.36**	0.713	0.735	0.719	0.701	0.715
new-thyroid1-5	**5.14**	0.932	0.946	0.948	0.911	0.95
new-thyroid2-5	**5.14**	0.927	0.964	0.945	0.927	0.945
ecoli2-5	**5.46**	0.756	0.778	0.816	0.8	0.804
segment0-5	**6.02**	0.963	0.972	0.982	0.977	0.97
glass6-5	**6.38**	0.875	0.809	0.809	0.809	0.826
yeast3-5	**8.1**	0.722	0.744	0.734	0.72	0.733
ecoli3-5	**8.6**	0.552	0.566	0.582	0.653	0.635
page-blocks0-5	**8.79**	0.552	0.593	0.591	0.612	0.617
Media		0.688	0.708	0.718	0.71	0.721

Tabla 20 Resultados de FARC-HD original y resultados de ensembles en clasificación no balanceada con Kappa

Al igual que con el AUC, en media todos los ensembles superan al algoritmo original, pero de nuevo mostramos los resultados del test de Wilcoxon comparando el original con el de tamaño 40 y voto ponderado.

Comparación	R+	R-	p-valor	Hipótesis α=0.05	Selección
FARC vs P40FARC	14.5	238.5	0.0004	Rechazado	**P40FARC**

Tabla 21 Wilcoxon mejor ensemble con FARC-HD original

De nuevo, con los resultados de la Tabla 21, se comprueba que el bagging de 40 clasificadores y voto ponderado mejora estadísticamente al FARC-HD original usando tanto Kappa como AUC como medida de rendimiento para el clasificador en conjuntos de datos no balanceados. No obstante, en las siguientes secciones se ha realizado el mismo estudio utilizando la media geométrica.

5.2.3 Media geométrica

A continuación, de la misma forma que en las secciones 5.2.1 y 5.2.2 en la tabla se muestran los resultados de los ensembles usando la media geométrica como medida de rendimiento.

Dataset	IR	Bagging			
		Voto simple		Voto ponderado	
		10	40	10	40
glass1-5	**1.82**	0.725	0.739	0.725	0.767
ecoli-0_vs_1-5	**1.86**	0.983	0.983	0.983	0.979
wisconsin-5	**1.86**	0.963	0.97	0.967	0.971
pima-5	**1.87**	0.754	0.762	0.74	0.763
iris0-5	**2**	1	1	1	1
glass0-5	**2.06**	0.815	0.833	0.824	0.833
yeast1-5	**2.46**	0.71	0.728	0.717	0.721
haberman-5	**2.78**	0.58	0.609	0.604	0.604
vehicle2-5	**2.88**	0.948	0.953	0.954	0.955
vehicle1-5	**2.9**	0.686	0.69	0.693	0.694
vehicle3-5	**2.99**	0.696	0.696	0.692	0.711
glass-0-1-2-3_vs_4-5-6-5	**3.2**	0.868	0.923	0.894	0.91
vehicle0-5	**3.25**	0.94	0.951	0.949	0.955
ecoli1-5	**3.36**	0.879	0.874	0.861	0.872
new-thyroid1-5	**5.14**	0.955	0.968	0.949	0.98
new-thyroid2-5	**5.14**	0.97	0.966	0.951	0.966
ecoli2-5	**5.46**	0.892	0.898	0.904	0.897
segment0-5	**6.02**	0.988	0.994	0.989	0.984
glass6-5	**6.38**	0.868	0.868	0.868	0.87
yeast3-5	**8.1**	0.919	0.928	0.92	0.925
ecoli3-5	**8.6**	0.814	0.829	0.858	0.868
page-blocks0-5	**8.79**	0.877	0.877	0.875	0.872
Media		0.856	0.865	0.86	0.868

Tabla 22 Resultados de ensembles en clasificación no balanceada con media geométrica

Al igual que con las medidas anteriores en media, el voto ponderado obtiene mejor resultado que el simple. A continuación se muestra la tabla resumen de los resultados anteriores.

Configuración	Datasets con mejor resultado	Datasets con peor resultado
Tamaño 10 voto simple	5	14
Tamaño 10 voto ponderado	2	6
Tamaño 40 voto simple	6	1
Tamaño 40 voto ponderado	14	3

Tabla 23 Resumen de los resultados con media geométrica

De nuevo, como en el estudio con el AUC y Kappa, en media el ensemble de tamaño 40 con voto ponderado es el que mejores resultados ofrece ya que tiene el mejor rendimiento en 14 datasets de 22.

A continuación, se mostrarán los resultados del test de Wilcoxon del mismo modo que se ha realizado con el Kappa y AUC.

5.2.3.1 Influencia del tipo de voto

En la siguiente tabla se muestran los resultados de los resultados del test de Wilcoxon para el estudio de la influencia del tipo de voto con diferentes tamaños

Comparación	R+	R-	p-valor	Hipótesis α=0.05	Selección
S10FARC vs P10FARC	85.5	167.5	0.1913	Aceptado	**P10FARC**
S40FARC vs P40FARC	118.5	135.5	0.8288	Aceptado	**P40FARC**

Tabla 24 Wilcoxon para el tipo de voto usando la media geométrica

Se puede concluir que no existen diferencias estadísticas entre los ensembles del mismo tamaño variando el tipo de voto. Este test, igual que con el Kappa y el AUC no ha sido determinante para poder decidir si un voto es mejor que el otro, pero siempre se favorece el mayor rango al voto ponderado. Por esta razón se seleccionan los ensembles del voto ponderado para el estudio posterior.

5.2.3.2 Influencia del tamaño del ensemble

Por no existir diferencias entre usar un tipo de voto u otro con ensembles del mismo tamaño y porque ofrecen mejor resultado en media, para este estudio del tamaño solo compararemos los que usan el voto ponderado.

Comparación	R+	R-	p-valor	Hipótesis α=0.05	Selección
P10FARC vs P40FARC	41.5	211.5	0.0057	Aceptado	**P40FARC**

Tabla 25 Wilcoxon para el tamaño según la Tabla 24

Los resultados del test, permiten decir que existen diferencias entre ambos clasificadores debido a que el p-valor es menor que $\alpha = 0.05$. Al igual que con las medidas anteriores, el ensemble de 40 clasificadores con voto ponderado supera estadísticamente a FARC-HD original.

5.2.3.3 Comparación con FARC-HD original

Del mismo modo que con el AUC y el Kappa, en la Tabla 26 se han añadido los resultados de FARC-HD original. Para realizar la comparativa se usará de nuevo el ensemble de 40 clasificadores con voto ponderado.

Dataset	IR	FARC	Bagging			
			Voto simple		Voto ponderado	
			10	40	10	40
glass1-5	**1.82**	0.729	0.725	0.739	0.725	0.767
ecoli-0_vs_1-5	**1.86**	0.98	0.983	0.983	0.983	0.979
wisconsin-5	**1.86**	0.968	0.963	0.97	0.967	0.971
pima-5	**1.87**	0.752	0.754	0.762	0.74	0.763
iris0-5	**2**	1	1	1	1	1
glass0-5	**2.06**	0.789	0.815	0.833	0.824	0.833
yeast1-5	**2.46**	0.718	0.71	0.728	0.717	0.721
haberman-5	**2.78**	0.578	0.58	0.609	0.604	0.604
vehicle2-5	**2.88**	0.931	0.948	0.953	0.954	0.955
vehicle1-5	**2.9**	0.685	0.686	0.69	0.693	0.694
vehicle3-5	**2.99**	0.705	0.696	0.696	0.692	0.711
glass-0-1-2-3_vs_4-5-6-5	**3.2**	0.862	0.868	0.923	0.894	0.91
vehicle0-5	**3.25**	0.932	0.94	0.951	0.949	0.955
ecoli1-5	**3.36**	0.874	0.879	0.874	0.861	0.872
new-thyroid1-5	**5.14**	0.965	0.955	0.968	0.949	0.98
new-thyroid2-5	**5.14**	0.952	0.97	0.966	0.951	0.966
ecoli2-5	**5.46**	0.88	0.892	0.898	0.904	0.897
segment0-5	**6.02**	0.983	0.988	0.994	0.989	0.984
glass6-5	**6.38**	0.909	0.868	0.868	0.868	0.87
yeast3-5	**8.1**	0.921	0.919	0.928	0.92	0.925
ecoli3-5	**8.6**	0.81	0.814	0.829	0.858	0.868
page-blocks0-5	**8.79**	0.871	0.877	0.877	0.875	0.872
Media		0.854	0.856	0.865	0.86	0.868

Tabla 26 Resultados de FARC-HD original y ensembles en clasificación no balanceada con la media geométrica

En la tabla anterior, al igual que sucedía con las anteriores medidas de rendimiento, se sigue observando que en media todos superan al algoritmo original, solo hay un caso en el que se empeora el rendimiento con todos los ensembles, esto ocurre en el dataset glass6-5. A parte de este detalle, ya que en las secciones anteriores se ha determinado el mejor ensemble es el de 40 clasificadores con voto ponderado a continuación se muestran los resultados del test de Wilcoxon comparando éste con FARC-HD original.

Comparación	R+	R-	p-valor	Hipótesis α=0.05	Selección
FARC vs P40FARC	25.5	227.5	0.0012	Rechazado	**P40FARC**

Tabla 27 Wilcoxon mejor ensemble con FARC-HD original

De los resultados de la Tabla 27 se puede concluir que el ensemble de tamaño 40, con voto ponderado supera el original, igual que se ha concluido con las otras medidas de rendimiento anteriores.

5.4 Comparación con otras técnicas de Bagging

En el estudio experimental anterior se ha concluido que la configuración que mas favorece para datasets no balanceados es con 40 FARC-HD y usando cada una de sus predicciones como voto ponderado. En esta sección se analizará el rendimiento de los ensembles de 40 clasificadores (con ambos tipos de voto) comparándolo con resultados de clasificadores obtenidos mediante diferentes técnicas de Bagging que el autor expone en [1].

En la Tabla 28 se muestra una breve descripción de los diferentes ensembles basados en bagging que usaremos para comparar los mejores resultados de nuestro estudio anterior. Debemos destacar que todos ellos están compuestos por 40 clasificadores base (árboles de decisión C4.5) para realizar una comparativa justa.

Ensembles con bagging		
Siglas	**Método**	**Descripción**
UB4	UnderBagging	Bagging con undersampling sobre la clase mayoritaria. Se aplica una selección aleatoria de ejemplos de la clase negativa.
UB24	UnderBagging2	Modificación de UnderBagging donde además se remuestrean con reemplazamiento los ejemplos de la clase positiva.
OB4	OverBagging	Bagging con oversamplig sobre la clase minoritaria. Se aplica una replicación aleatoria de ejemplos de la clase positiva.
OB24	OverBagging2	Modificación de OverBagging donde además se remuestrean con reemplazamiento los ejemplos de la clase negativa.
UOB4	UnderOverBagging	Técnica que aplica tanto UnderBagging como OverBagging. El número de ejemplos utilizados de cada clase se incrementa en cada bag.
SBAG4	SMOTEBagging	Bagging con oversamplig sobre la clase minoritaria. Se aplica SMOTE para obtener ejemplos de la clase positiva. El número de ejemplos de cada clase también se incrementa en cada bag.

Tabla 28 Descripción métodos de Bagging [1]

A continuación en la Tabla 29 se muestran los resultados de diferentes técnicas de Bagging expuestas en [1] con 40 clasificadores. Las columnas correspondientes a las siglas S40 y P40 corresponden a los resultados de los ensembles con voto simple y ponderado respectivamente.

Data-set	IR	Ensembles con Bagging						Bagging FARC	
		UB4	UB24	OB4	OB24	UOB4	SBAG4	S40	P40
glass1	1.82	0.737	0.752	0.758	0.78	0.774	0.728	0.746	0.771
ecoli-0_vs_1	1.86	0.980	0.980	0.980	0.980	0.980	0.983	0.983	0.980
wisconsin	1.86	0.960	0.971	0.964	0.973	0.961	0.960	0.970	0.971
pima	1.9	0.760	0.753	0.715	0.738	0.736	0.751	0.762	0.763
iris0	2	0.990	0.980	0.980	0.980	0.980	0.980	1.000	1.000
glass0	2.06	0.814	0.824	0.827	0.834	0.838	0.839	0.834	0.834
yeast1	2.46	0.722	0.721	0.717	0.734	0.723	0.734	0.729	0.723
vehicle1	2.52	0.787	0.761	0.716	0.757	0.724	0.769	0.702	0.705
vehicle2	2.52	0.963	0.964	0.957	0.968	0.965	0.966	0.954	0.955
vehicle3	2.52	0.802	0.784	0.706	0.738	0.722	0.763	0.704	0.716
haberman	2.68	0.664	0.668	0.614	0.642	0.588	0.656	0.632	0.632
glass-0-1-2-3_vs_4-5-6	3.19	0.904	0.917	0.935	0.926	0.912	0.945	0.926	0.912
vehicle0	3.23	0.952	0.954	0.945	0.966	0.952	0.965	0.951	0.955
ecoli1	3.36	0.900	0.902	0.865	0.888	0.886	0.900	0.878	0.876
new-thyroid2	4.92	0.958	0.938	0.932	0.932	0.954	0.961	0.969	0.969
new-thyroid1	5.14	0.963	0.969	0.940	0.963	0.977	0.975	0.969	0.980
ecoli2	5.46	0.884	0.881	0.892	0.897	0.880	0.888	0.903	0.901
segment0	6.01	0.988	0.986	0.992	0.993	0.991	0.994	0.994	0.984
glass6	6.38	0.904	0.926	0.886	0.914	0.917	0.931	0.878	0.881
yeast3	8.11	0.934	0.944	0.906	0.911	0.917	0.944	0.928	0.925
ecoli3	8.19	0.908	0.894	0.740	0.810	0.811	0.885	0.851	0.881
page-blocks0	8.77	0.958	0.959	0.938	0.944	0.953	0.953	0.878	0.873
Media		0.883	0.883	0.859	0.876	0.870	0.885	0.870	0.872

Tabla 29 Resultados de otras técnicas de Bagging

Donde las siglas de las columnas del apartado Bagging basados en ensembles, corresponden a diferentes modalidades de Bagging expuestas en la siguiente tabla:

Para la comparación de clasificadores hemos aplicado en este caso la prueba de rangos alineados de Friedman la cual se ha explicado en la Sección 4.3, la cual asocia el menor rango al mejor clasificador, y a este clasificador se le denomina método de control (MC).De esta forma podremos determinar si existen diferencias entre las técnicas propuestas por el autor en [1] y nuestros ensembles de tamaño 40. Se realizaran diferentes test para cada tipo de voto por separado.

5.2.4.1 Voto simple con 40 clasificadores

En la Tabla 30 se muestran os resultados del test de rangos alineados de Friedman.

Técnica Bagging	UB4	UB2	OB4	OB2	UOB4	SBAG4	S40FARC
Rangos Alin Friedman	68.0681	64.0681	112.9318	74.6818	91.4318	47.4545	83.8636
APV Holm	0.2505	0.2505	0.0001	0.1286	0.0053	MC	0.0271

Tabla 30 Resultados Friedman para S40FARC

El p-valor obtenido en el test anterior es 0.0032, que al ser menor que 0.05 podemos decir que existen diferencias significativas entre los clasificadores, siendo el mejor el de menor rango, es decir SBAG4, el método de control, el cual mejora estadísticamente a nuestra propuesta. Entre los 7 clasificadores comparados, el S40FARC ocuparía el 5 puesto. En la siguiente ilustración se muestra una gráfica con los rangos obtenidos en el test por cada clasificador.

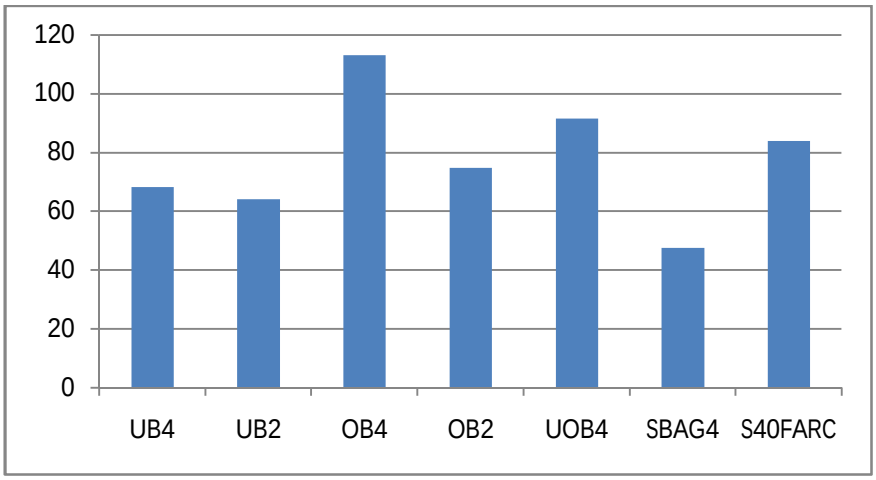

Gráfico 2 Rangos alineados de Friedman para cada técnica

No obstante se han realizado las pruebas de Wilcoxon comparando el S40FARC con el resto de clasificadores obteniendo los siguientes resultados

Comparación	R+	R-	p-valor	Hipótesis α=0.05	Selección
S40FARC vs UB4	105	148	0.4851	Aceptado	**UB4**
S40FARC vs UB2	87	166	0.1997	Aceptado	**UB2**
S40FARC vs OB4	185	68	0.0575	Aceptado	**S40FARC**
S40FARC vs OB2	103	150	0.4454	Aceptado	**OB2**
S40FARC vs UOB4	137	116	0.7330	Aceptado	**S40FARC**
S40FARC vs SBAG4	60	193	0.0308	Rechazado	**SBAG4**

Tabla 31 Wilcoxon de S40FARC con el varias técnicas de [1]

Analizando los resultados anteriores podemos concluir que no existen diferencias estadísticas entre S40FARC y el resto de clasificadores excepto en el caso de SBAG que lo supera significativamente. A pesar de esto podemos decir que obtiene el mayor rango al compararlo con OB4 y UOB4 (aunque no existan diferencias estadísticamente entre ellos), al

igual que ocurre con los rangos alineados de Friedman mostrados anteriormente donde S40FARC obtenía menor rango que OB4 y UOB4.

Con el voto ponderado según los estudios de las secciones anteriores se espera que se obtengan mejores.

5.2.4.1 Voto ponderado con 40 clasificadores

Del mismo modo que con el voto simple, a continuación mostramos los resultados de la prueba de los rangos alineados de Friedman comparando el P40FARC con los clasificadores usados para la comparación anterior con el voto simple.

Técnica Bagging	**UB4**	**UB2**	**OB4**	**OB2**	**UOB4**	**SBAG4**	**P40FARC**
Rangos Alin Friedman	68.0909	64.4999	112.7499	75.2045	91.8181	48.5454	81.5909
APV Holm	0.2921	0.2921	1.0810	0.1422	0.0064	MC	0.0559

Tabla 32 Resultados Friedman para P40FARC

En la tabla anterior se muestran os resultados del test de rangos alineados de Friedman y e l p-valor obtenido, de nuevo es 0.0032. Como este valor es menor que 0.05 podemos decir que existen diferencias significativas entre los clasificadores, siendo el mejor el de menor rango, es decir, en este caso de nuevo es SBAG4. Entre los 7 clasificadores comparados, el P40FARC ocuparía el 5 puesto al igual que sucedía con S40FARC. En la siguiente ilustración se muestra una gráfica con los rangos obtenidos en el test por cada clasificador.

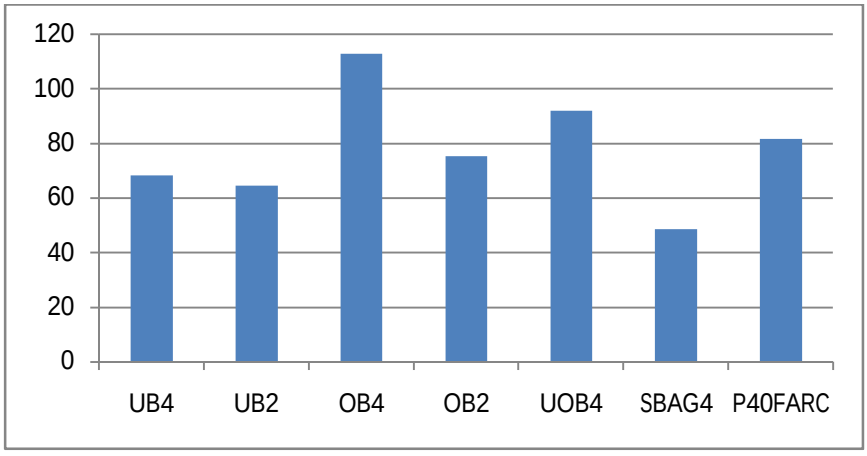

Gráfico 3 Rangos alineados de Friedman para cada técnica

El test de rangos alineados de Friedman, de nuevo lo único que nos permite afirmar es que existen diferencias significativas entre el SBAG4 y el resto, siendo éste el mejor. Por esta razón hemos realizado las comparativas de Wilcoxon de P40FARC con el resto de clasificadores. Los resultados de estas pruebas se muestran a continuación.

Comparación	R+	R-	p-valor	Hipótesis α=0.05	Selección
P40FARC vs UB4	105.5	147.5	0.4549	Aceptado	**UB4**
P40FARC vs UB2	95.5	157.5	0.2891	Aceptado	**UB2**
P40FARC vs OB4	189	64	0.0424	Rechazado	**P40FARC**
P40FARC vs OB2	104.5	148.5	0.5202	Aceptado	**OB2**
P40FARC vs UOB4	139.5	113.5	0.6639	Aceptado	**P40FARC**
P40FARC vs SBAG4	73	180	0.0824	Aceptado	**SBAG4**

Tabla 33 Wilcoxon de S40FARC con el varias técnicas de [1]

En el apartado del voto simple también obtenía mejor rango el S40FARC que OB4 y UOB4 pero el Wilcoxon no nos ha permitido decir que existieran diferencias. En esta ocasión, el test de Wilcoxon nos permite afirmar que P40FARC supera estadísticamente al OB4 (que este es el que peores resultados ofrece inicialmente), no se supera estadísticamente a UOB4 pero seguimos consiguiendo mayor rango.

Además en este caso, usando el voto ponderado, SBAG4 ya no nos supera estadísticamente (con el voto simple si nos superaba) ya que el p-valor correspondiente es superior a 0.05. Por lo que vemos que a pesar de la sencillez de la propuesta, el ensemble de 40 clasificadores con voto ponderado se muestra competitivo con respecto a los clasificadores del estado del arte.

6. Estudio experimental: análisis de los resultados de la propuesta preliminar de generación de un clasificador difuso

Como se ha explicado en la Sección 3.2 la última parte de este proyecto está dedicada a desarrollar un clasificador difuso obtenido a partir del conjunto de clasificadores que forman el ensemble. Mejorar un SCBRDs con bagging implica la creación de un conjunto de bases de reglas y bases de datos asociadas a cada clasificador que formará parte del ensemble, esta propuesta consiste en reducir el número de reglas eliminando redundancias para formar un nuevo clasificador. Hemos creado 4 clasificadores a partir de las 4 configuraciones estudiadas en las secciones anteriores (ensembles de 10 y 40 clasificadores con voto simple y ponderado).

En las secciones 6.1 y 6.2 se muestran los resultados para clasificación estándar y no balanceada respectivamente. Resultados que han sido obtenidos en esta última propuesta de *reducción de reglas* con el objetivo de determinar si resulta rentable dicha reducción.

6.1 Resultados en clasificación estándar

Del mismo modo que se han presentado los resultados en la Sección 5.1, la siguiente tabla muestra los resultados de FARC-HD original junto a los resultados de los clasificadores obtenidos tras la reducción de reglas del ensemble medidos con el accuracy.

Dataset	FARC	SCBRDs bagging			
		Voto simple		Voto ponderado	
		10	40	10	40
automobile	73.68	57.51	58.47	53.35	60.75
balance	87.53	87.2	87.85	83.7	86.57
cleveland	58.89	59.29	56.23	59.58	58.5
contraceptive	53.97	49.83	48.87	49.55	50.56
ecoli	84.6	75.98	72.65	76.28	77.38
glass	65.02	60.01	53.64	58.68	59.34
hayes-roth	78.8	72.57	68.09	73.09	73.14
iris	94.67	95.33	96	96	95.5
newthyroid	93.95	89.77	84.65	89.3	89.42
page-blocks	95.27	91.98	91.54	92.78	92.89
segment	92.9	83.07	80.26	81.86	84.52
shuttle	97.48	85.44	85.38	88.4	89.17
tae	57.84	48.92	45.71	50.54	50.75
thyroid	94.26	93.03	89.77	91.03	92.02
vehicle	68.45	58.75	61.48	60.4	62.27
vowel	74.95	52.32	47.88	50.51	56.41
wine	95.54	94.39	95.54	95.55	95.26
yeast	59.65	52.51	51.62	51.95	53.93
Media	79.3	72.66	70.87	72.36	73.8

Tabla 34 Resultados de FARC-HD original y de la reducción de reglas en clasificación estándar

La Tabla 34 muestra claras deficiencias en los SCBRDs generados a partir del bagging. El que mejor resultado ofrece es la reducción de reglas del P40FARC pero no es suficiente como para superar al algoritmo original. La mejor tasa de acierto entre los nuevos clasificadores es 73.80% y FARC-HD original consigue un 79.3%, luego se ha reducido el accuracy un 5.5%. Por esta razón, decimos que esta propuesta no ha resultado ser eficiente ya que empeoramos los resultados originales, pero es únicamente una propuesta preliminar.

Se aprecia en ciertos dataset que algunos modelos si consigue mejorar al original, pero en media general no. Si existiese un sistema encargado de clasificar los datos del iris que utilice el FARC-HD original, sería interesante generar un ensemble con bagging y aplicar la reducción de reglas propuestas para conseguir una mejora.

En la Sección 6.2 se muestran los resultados de estos clasificadores pero con conjuntos de datos no balanceados de la misma forma que se han presentado los resultados del estudio experimental del Bagging.

6.2 Resultados en clasificación no balanceada

A pesar de no ofrecer buenos resultados en clasificación estándar también se analizarán los resultados en clasificación no balanceada con las mismas 3 medidas de rendimiento.

6.2.1 AUC

Dataset	IR	FARC	SCBRDs bagging			
			Voto simple		Voto ponderado	
			10	40	10	40
glass1-5	**1.82**	0.739	0.565	0.547	0.613	0.588
ecoli-0_vs_1-5	**1.86**	0.98	0.947	0.953	0.963	0.921
wisconsin-5	**1.86**	0.969	0.955	0.957	0.965	0.948
pima-5	**1.9**	0.753	0.741	0.721	0.711	0.729
iris0-5	**2**	1	1	1	1	1
glass0-5	**2.06**	0.792	0.717	0.612	0.724	0.638
yeast1-5	**2.46**	0.72	0.668	0.675	0.688	0.682
vehicle1-5	**2.52**	0.694	0.629	0.587	0.601	0.578
vehicle2-5	**2.52**	0.932	0.855	0.721	0.837	0.704
vehicle3-5	**2.52**	0.709	0.626	0.564	0.576	0.557
haberman-5	**2.68**	0.604	0.534	0.548	0.576	0.519
glass-0-1-2-3_vs_4-5-6-5	**3.19**	0.87	0.836	0.801	0.848	0.784
vehicle0-5	**3.23**	0.933	0.733	0.712	0.829	0.776
ecoli1-5	**3.36**	0.876	0.767	0.725	0.774	0.707
new-thyroid2-5	**4.92**	0.954	0.9	0.886	0.914	0.886
new-thyroid1-5	**5.14**	0.966	0.929	0.886	0.906	0.886
ecoli2-5	**5.46**	0.884	0.642	0.51	0.591	0.508
segment0-5	**6.01**	0.983	0.85	0.822	0.874	0.782
glass6-5	**6.38**	0.914	0.859	0.764	0.795	0.831
yeast3-5	**8.11**	0.921	0.867	0.816	0.882	0.836
ecoli3-5	**8.19**	0.819	0.59	0.556	0.651	0.568
page-blocks0-5	**8.77**	0.871	0.848	0.838	0.828	0.842
Media		0.858	0.775	0.736	0.779	0.739

Tabla 35 Resultados de FARC-HD original y de la reducción de reglas en clasificación no balanceada con AUC

Los resultados de la Tabla 35 muestran lo mismo que en clasificación estándar, tampoco se mejora el rendimiento de FARC-HD original. Únicamente en el dataset iris-5 se consigue igualar el rendimiento de algoritmo original. En nuestro estudio no es suficiente que el nuevo modelo de clasificación iguale el rendimiento ya que para conseguir el nuevo clasificador se sigue un largo proceso (creación del ensemble mas reducción de reglas), el cual si no consigue mejorar el rendimiento original, no es necesario que lo apliquemos. Nos quedaríamos con el FARC-HD original.

6.2.2 Kappa

En la sección anterior, con los resultados del AUC se hemos concluido que la propuesta no funciona correctamente, la tabla siguiente muestra los resultados de los mismos clasificadores pero con el Kappa.

Dataset	IR	FARC	SCBRDs bagging			
			Voto simple		Voto ponderado	
			10	40	10	40
glass1-5	**1.82**	0.459	0.145	0.112	0.228	0.203
ecoli-0_vs_1-5	**1.86**	0.96	0.908	0.918	0.938	0.87
wisconsin-5	**1.86**	0.93	0.916	0.919	0.929	0.903
pima-5	**1.9**	0.485	0.457	0.455	0.423	0.467
iris0-5	**2**	1	1	1	1	1
glass0-5	**2.06**	0.573	0.435	0.247	0.447	0.286
yeast1-5	**2.46**	0.401	0.361	0.379	0.374	0.4
vehicle1-5	**2.52**	0.358	0.267	0.224	0.219	0.2
vehicle2-5	**2.52**	0.828	0.755	0.518	0.722	0.484
vehicle3-5	**2.52**	0.365	0.283	0.175	0.189	0.15
haberman-5	**2.68**	0.2	0.077	0.123	0.182	0.052
glass-0-1-2-3_vs_4-5-6-5	**3.19**	0.758	0.705	0.678	0.731	0.652
vehicle0-5	**3.23**	0.819	0.531	0.507	0.682	0.613
ecoli1-5	**3.36**	0.713	0.576	0.511	0.588	0.478
new-thyroid2-5	**4.92**	0.927	0.863	0.831	0.881	0.845
new-thyroid1-5	**5.14**	0.932	0.905	0.84	0.841	0.837
ecoli2-5	**5.46**	0.756	0.361	0.032	0.239	0.025
segment0-5	**6.01**	0.963	0.746	0.73	0.812	0.643
glass6-5	**6.38**	0.875	0.765	0.567	0.622	0.745
yeast3-5	**8.11**	0.722	0.706	0.675	0.692	0.689
ecoli3-5	**8.19**	0.552	0.178	0.156	0.38	0.192
page-blocks0-5	**8.77**	0.552	0.59	0.602	0.578	0.601
Media		0.688	0.57	0.509	0.577	0.515

Tabla 36 Resultados de FARC-HD original y de la reducción de reglas en clasificación no balanceada con Kappa

Observando los resultados del Kappa en la Tabla 36 se aprecia que en media también se empeoran los resultados. En el caso del dataset iris0-5 se sigue manteniendo el rendimiento, pero además en page-blocks0-5 consigue que los nuevos clasificadores superen la cifra del FARC-HD original.

6.2.3 Media geométrica

Por último para completar este estudio se presentan en la siguiente tabla los resultados de FARC-HD original y los nuevos clasificadores obtenidos con la media geométrica.

Dataset	IR	FARC	Reducción bagging			
			Voto simple		Voto ponderado	
			10	40	10	40
glass1-5	**1.82**	0.729	0.424	0.364	0.461	0.472
ecoli-0_vs_1-5	**1.86**	0.98	0.946	0.952	0.962	0.913
wisconsin-5	**1.86**	0.968	0.955	0.957	0.964	0.947
pima-5	**1.9**	0.752	0.737	0.706	0.701	0.713
iris0-5	**2**	1	1	1	1	1
glass0-5	**2.06**	0.789	0.68	0.419	0.674	0.512
yeast1-5	**2.46**	0.718	0.63	0.632	0.663	0.638
vehicle1-5	**2.52**	0.685	0.564	0.442	0.514	0.418
vehicle2-5	**2.52**	0.931	0.842	0.638	0.819	0.615
vehicle3-5	**2.52**	0.705	0.519	0.357	0.397	0.288
haberman-5	**2.68**	0.578	0.363	0.37	0.47	0.252
glass-0-1-2-3_vs_4-5-6-5	**3.19**	0.862	0.81	0.768	0.821	0.742
vehicle0-5	**3.23**	0.932	0.683	0.653	0.815	0.746
ecoli1-5	**3.36**	0.874	0.744	0.648	0.731	0.641
new-thyroid2-5	**4.92**	0.952	0.891	0.867	0.905	0.876
new-thyroid1-5	**5.14**	0.965	0.923	0.873	0.899	0.87
ecoli2-5	**5.46**	0.88	0.475	0.063	0.33	0.063
segment0-5	**6.01**	0.983	0.824	0.797	0.854	0.719
glass6-5	**6.38**	0.909	0.844	0.625	0.682	0.807
yeast3-5	**8.11**	0.921	0.86	0.799	0.877	0.82
ecoli3-5	**8.19**	0.81	0.268	0.213	0.493	0.288
page-blocks0-5	**8.77**	0.871	0.843	0.831	0.821	0.836
Media		0.854	0.719	0.635	0.721	0.644

Tabla 37 Resultados de FARC-HD original y de la reducción de reglas en clasificación no balanceada con la media geométrica

En esta ocasión, al igual que con el AUC lo único que se consigue es mantener el rendimiento de FARC-HD original en el dataset iris, en los demás se empeora, por lo que los resultados en media tampoco superan al original.

7. Conclusiones

Después del estudio realizado a lo largo de las seccione s 5.1 y 5.2 se ha llegado a las siguientes conclusiones:

En clasificación estándar es posible mejorar la tasa de acierto del algoritmo FARC-HD utilizando Bagging, que cualquiera de las configuraciones para los ensembles superan estadísticamente al algoritmo original, y el que mejores resultados ofrece es utilizando 40 clasificadores y el voto simple, S40FARC.

Para la clasificación no balanceada se puede afirmar que la configuración más apropiada entre las estudiadas es la de 40 con voto ponderado. Esta afirmación queda fundamentada con los estudios realizados en las secciones 5.2.1, 5.2.2 y 5.2.3 con el AUC, Kappa y media geométrica respectivamente, donde cada medida nos ha permitido llegar a las siguientes conclusiones.

El estudio con el AUC es el que nos ha llevado a afirmar que existen diferencias muy significativas entre los resultados del algoritmo original y el Bagging con 40 clasificadores usando el voto ponderado, P40FARC.

En la Tabla 17 de resumen de resultados con Kappa se aprecian 2 mejoras, al pasar de voto simple a ponderado y al aumentar el tamaño de clasificadores de 10 a 40, siendo mucho más significativa esta segunda. En ambos casos se consigue aumentar el número de datasets donde obtiene el mayor rendimiento y disminuir el número en los que obtiene el peor rendimiento. Además también se ha concluido con que el P40FARC mejora al algoritmo original igual que con el AUC.

El análisis con la media geométrica ha concluido que los ensembles de tamaño 40 superan estadísticamente al original, en cambio los de tamaño 10 no. Ofreciendo mejores resultados el de voto ponderado.

En cuanto a la técnica estudiada en la Sección 6, los resultados de los nuevos SCBRDs obtenidos tras la reducción de reglas del bagging, no ha conseguido mejorar el rendimiento del algoritmo FARCD-HD en ninguno de los dos tipos de clasificación. En clasificación estándar se ha reducido el porcentaje de acierto medio en un 5.5% en el mejor caso. En clasificación no balanceada los resultados en media tampoco superan al original pero en concreto, si que existen determinados datasets en los que se consigue igualar el rendimiento o superarlo dependiendo de la medida de rendimiento utilizada.

Con el objetivo de dar un por qué, analizando el proceso que se sigue en la reducción de reglas podemos intuir que las deficiencias que presentan los nuevos clasificadores pueden ser por varios motivos:

- La tercera fase de la reducción de reglas consiste en seleccionar de todo el conjunto de reglas del ensemble (no repetidas), aquellas reglas que son tan generales que no están incluidas en otras reglas. Es decir, cuando una pareja de reglas tienen parte del antecedente común y una con el antecedente más largo que la otra, se selecciona para el nuevo clasificador la regla con el antecedente más corto (la más generalista). Esto puede provocar que el clasificador generalice demasiado y no existan reglas suficientemente específicas para las clases que lo necesiten. En datasets con alto solapamiento de clases el clasificador tenderá a predecir la clase mayoritaria entre los ejemplos solapados ya que usará más cantidad de reglas generalistas que específicas.
- La elección de la base de datos que definirá las funciones de pertenencia para el nuevo clasificador, se realiza evaluando cada clasificador que forma el ensemble sobre el conjunto de datos de entrenamiento original (sin muestrear) y seleccionamos la base de datos que mayor rendimiento ofrezca sobre dicho conjunto. Esto puede dar problemas debido a que las funciones de pertenencia no están ajustadas a las reglas que forman el nuevo clasificador. Este desajuste es debido a que en el nuevo clasificador, ciertas reglas se empleen en la clasificación sobre una base de datos diferente a la que usaron para su aprendizaje. Este problema se muestra con el ejemplo siguiente:

  Partimos de un ensemble ya creado de 5 clasificadores: A, B, C, D y E. Cada clasificador tiene 5, 6, 3, 8 y 10 reglas respectivamente y cada uno su base de datos (todas muy parecidas entre sí, pero diferentes). Suponiendo que tras la fase de reducción de reglas, nuestro nuevo clasificador está formado por una regla de cada clasificador base, y que la base de datos más precisa en el conjunto de datos original es la del clasificador A. Nuestro nuevo clasificador usaría 4 de sus reglas apoyándose en una base de datos diferente a la que se usó para el aprendizaje de estas.

Las causas expuestas anteriormente, nos llevan a abrir nuevas líneas futuras de estudio dedicadas la creación de técnicas de generación de clasificadores a partir del conjunto de bases de reglas y de datos que forman los ensembles.

8. Líneas futuras

En el estudio experimental de la Sección 5 se ha conseguido mejorar FARC-HD original con técnicas de bagging para la creación de ensembles, pero en el estudio de la Sección 6, donde se ha intentado reducir el conjunto de clasificadores que forman el ensemble para formar un único clasificador se ha comprobado que no se supera a FARC-HD ofreciendo peores resultados que este y por supuesto peores que el ensemble.

Por este motivo, a modo de abrir futuras líneas de investigación, en las secciones 8.1, 8.2 y 8.3 se proponen nuevas propuestas para la creación del nuevo clasificador teniendo en cuenta las deficiencias expuestas la Sección 7.

8.1 Selección de reglas más frecuentes

Los clasificadores que se aprenden con el bagging para la creación de un ensemble como ya se ha explicado a lo largo del trabajo se entrenan con muestras *bootstrap* del conjunto de datos original. Por esto las reglas de los clasificadores serán muy parecidas o incluso las mismas. Por este motivo, en lugar de eliminar las reglas repetidas y reglas demasiado específicas (ya incluidas por otras más generales), podemos seguir otra estrategia midiendo el porcentaje o frecuencia de aparición de cada regla.

Es decir, al existir varios clasificadores, aparecerán reglas las cuales estén presentes en varios clasificadores. Esta idea sigue la línea de seleccionar las reglas más frecuentes que aparezcan en el conjunto de clasificadores.

Para plantear esta idea, se han realizado unas pruebas preliminares en las que se ha detectado un problema evidente: Es necesario fijar un umbral (porcentaje de aparición de reglas) que determinará que reglas son seleccionadas para el nuevo clasificador. Un umbral muy alto puede suponer que se seleccionen reglas muy buenas (las cuales aparecerán en la mayoría de clasificadores) y precisas pero también nos interesará que se seleccionen reglas especificas de cada clasificador (ya que cada uno de estos tendrá reglas especificas para diferentes tipos de ejemplos a consecuencia del muestreo). Además a cada dataset le favorecerá un umbral diferente por lo que sería necesario establecer un método para fijar este umbral y que no sea fijado a mano.

Destacar que para esta técnica sería conveniente tener un tamaño de ensemble lo suficientemente grande con el objetivo de que se genere un número significativo de repeticiones de las reglas en los clasificadores que formen el ensemble.

8.2 Selección de reglas más específicas

Otro posible método para construir el nuevo clasificador que se podría emplear es seguir una técnica similar a la utilizada en la Sección 6 (unir todas bases de reglas de los clasificadores que forman el ensemble y eliminar redundancias) pero siguiendo el criterio inverso en la selección de reglas ante la aparición de redundancias. Es decir, en la propuesta desarrollada, cuando existían redundancias, las reglas seleccionadas para el nuevo clasificador son las más generalistas pero los resultados obtenidos no son buenos.

Por esto y debido a que la propuesta está afectando especialmente al rendimiento en conjuntos de datos no balanceados, proponemos que una posible continuación de este proyecto fuese realizar el estudio que hemos realizado pero en la fase de eliminación de reglas redundantes, las reglas escogidas para el nuevo clasificador fuesen las más específicas en lugar de los más generales. De esta manera tratará de mantener en la nueva base de reglas aquellas más específicas, lo que hipotéticamente beneficiará la predicción de los ejemplos de la clase minoritaria.

8.3 Técnicas de selección y fusión de las bases de datos

La técnica expuesta en la Sección 3.2.1 tiene un claro defecto en la selección de la base de datos, debido a que se escoge la base de datos que mejor rendimiento ofrece en el conjunto de entrenamiento original pero no tiene cuenta el resto de bases de datos del clasificador. Esto puede suponer un problema en el nuevo clasificador ya que varias de las reglas que lo forman se han podido entrenar con diferente base de datos a la seleccionada. Por esto, se puede intuir que las reglas pierdan efectividad.

Otras posibles técnicas para obtener una base de datos para nuestro nuevo clasificador que tenga en cuenta el resto de bases de datos podrían ser: realizar una media aritmética de todas las bases de datos componente a componente, utilizar algoritmos genéticos para reajustar las funciones de pertenencia, incluso se podría usar el algoritmo genético que usa FARC-HD en su última fase del aprendizaje para esto.

9. Bibliografía

[1] Galar, Mikel at al., "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches", *IEEE Transaction on systems, man and cybernetics – Part C: applications and Reviews,* vol. 42, num. 4, JULY 2012.

[2] Richard O. Duda, Peter E. Hart, David G. Stork. Wiley-Interscience. *Pattern Classification.* 653 p. Toronto: John Wiley & Sons, 2001. ISBN: 978-0471056690

[3] Q. Yang and X. Wu, "10 challenging problems in data mining research," International Journal of Information Technology and Decision Making, vol. 5, no. 4, pp. 597–604, 2006.

[4] 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.

[5] XindongWu, VipinKumar. "The Top Ten Algorithms in Data Mining". London: CRC, 2009.

[6] Jesús Alcalá-Fdez, Rafael Alcalá, and Francisco Herrera. "A Fuzzy Association Rule-Based Classification Model for High-Dimensional Problems with Genetic Rule Selection and Lateral Tuning".*IEEE Transactionson Fuzzy Systems*, 2011, vol. 19, pp. 857-872.

[7] Elkano Ilintxeta, Mikel. (2014). "IVOVO: sistema de multi-clasificación basado en datos intervalo-valorados" (Trabajo fin de grado). Universidad Pública de Navarra, Pamplona.

[8] Breiman, Leo. "Bagging Predictors". Berkeley: University of California, 1994. Technical Report No. 421.