Medidas difusas para la combinación de ensembles: una primera aproximación utilizando el rendimiento en conjuntos de datos altamente desequilibrados

M. Uriz, D. Paternain, H. Bustince, M. Galar
Departamento de Estadística, Informática y Matemáticas, *Universidad Pública de Navarra*,
Campus Arrosadía s/n, 31006 Pamplona, España
{mikelxabier.uriz, daniel.paternain, bustince, mikel.galar}@unavarra.es

***Resumen*—En este trabajo estudiamos la posibilidad de aprender medidas difusas basándonos en el rendimiento de los clasificadores para mejorar las funciones de agregación tradicionales utilizadas en los ensembles de clasificadores. Las medidas difusas son funciones çonjunto-valor", no necesariamente aditivas, y son la base para la construcción de las integrales difusas no lineales, como las integrales de Choquet y de Sugeno.Nuestra hipótesis es que teniendo en cuenta las interacciones entre los miembros del ensemble podemos llegar a un mejor rendimiento. Precisamente, las medidas difusas nos permiten modelar dichas interacciones. Nuestra propuesta consiste en obtener la medida difusa directamente de los datos considerando el rendimiento de cada subconjunto de clasificadores del ensemble. De esta manera, no necesitamos incluir otros métodos de aprendizaje para la medida difusa que nos pueden llevar a un sobre-ajuste. Para verificar la utilidad de la medida propuesta, consideraremos 33 conjuntos de datos altamente desequilibrados y desarrollaremos un estudio experimental completo comparando nuestra propuesta con otros métodos considerados comúnmente en la literatura.**

***Index Terms*—Agregaciones, Medidas Difusas, Clasificación, Ensembles, Datasets no balanceados**

I. Introducción

Los ensembles de clasificadores [1] son conocidos por ser una buena alternativa para mejorar el rendimiento de un único clasificador utilizando una combinación de varios clasificadores. En los ensembles, se espera que los clasificadores individuales que forman el ensemble tengan diversidad, de tal manera que su combinación nos lleve a una mejora de los resultados. Sin la diversidad, todos los clasificadores llegarían a las mismas respuestas y por lo tanto, su combinación sería la misma que si se considerarían por separado.

A la hora de construir un ensemble para un problema especifico, hay que tener en cuenta dos aspectos importantes: 1) cómo generar diversidad entre clasificadores; 2) cómo combinar la decisión de todos los clasificadores en una única salida. Aunque ambos aspectos son importantes para el rendimiento final del ensemble, el primero ha recibido más atención en la literatura y se han desarrollado varias formas de conseguir la diversidad, siendo Bagging [2] y Boosting [3] los más populares. Esto es debido a que se esperan mayores mejoras cambiando los clasificadores subyacentes que combinando sus salidas de diferentes formas. Sin embargo, no se puede pasar por alto la importancia de cómo se fusionan las salidas. Aunque la ganancia en rendimiento puede ser menor, se puede utilizar el conocimiento adquirido por cada clasificador para llevar a cabo una mejor combinación que simplemente utilizar la media aritmética.

En este articulo pretendemos mejorar el rendimiento de los ensembles solo modificando la fase de combinación. Obviamente, se han desarrollado varios métodos con este propósito, p.e., el voto ponderado, la combinación de Naive Bayes o los Decision Templates entre otros [1]. A diferencia de estos, nosotros nos centraremos en entender cómo colabora cada clasificador con el resto. Es decir, intentaremos modelar las coaliciones o intersecciones (positivas o negativas) que pueden formar dependiendo de su rendimiento. Para modelar este comportamiento, consideramos el uso de las medidas difusas y las integrales difusas [4], [5]. Debemos hacer notar que aunque estas han sido utilizadas anteriormente para la combinación de clasificadores, todo su potencial no ha sido explotado, dada la limitación impuesta es su construcción [6], [7].

Las dos funciones de agregación basadas en medidas difusas mas utilizadas son las integrales de Choquet [8] y Sugeno [9]. Ambas permiten agregar o fusionar información cuantitativa tendiendo en cuenta la interacción entre los datos mediante medias de la medida difusa asociada. De esta manera, esta familia de funciones de agregación pueden ser más potentes que el resto de agregaciones más comunes, como la media ponderada o los operadores OWA.

Por consiguiente, nuestra propuesta consiste en el uso de integrales difusas conjuntamente con medidas difusas para realizar la combinación de clasificadores. Para ello proponemos una nueva forma de obtener las medidas difusas a partir del rendimiento de cada clasificador, el cual refleja mejor como interaccionan realmente.

A fin de estudiar empíricamente la utilidad de la construcción de la medida difusa propuesta, hemos considerado el problema de las clases no balanceadas y los ensembles específicamente diseñados para atacar este problema [10]. Más concretamente, nos centramos en UnderBagging [11]. El estudio experimental se ha realizado sobre los 33 conjuntos de datos con mayor ratio de desequilibrio (no balanceo) del repositorio de KEEL [12]. Compararemos nuestra propuesta con el uso de otras funciones de agregación existentes para la combinación de clasificadores como las agregaciones ponderadas o las medidas difusas lambda [6], [7].

El resto del articulo esta organizado de la siguiente manera. En la Sección II, recordamos la teoría de las funciones de agregación necesaria para comprender el resto del trabajo. Después, la Sección III introduce los ensembles, las diferentes maneras de combinar clasificadores y las soluciones adaptadas para el problema del no balanceo. En la Sección IV, describimos cómo establecer los parámetros de las agregaciones ponderadas y también nuestra propuesta para la creación de la medida difusa. Las Secciones V y VI, explican el marco de trabajo experimental considerado para este estudio y analizan los resultados obtenidos. Finalmente, la sección VII concluye este trabajo.

II. Funciones de agregación

En esta sección recordamos varios conceptos sobre las funciones de agregación que van a ser usadas para agregar las salidas de los ensembles. Los operadores de agregación o las funciones de agregación son una herramienta matemática que se han vuelto esencial para los problemas de fusión de información. En la literatura es muy fácil encontrar monografías completas que cubren varias familias de funciones de agregación, como las t-normas, t-conormas, medias, integrales, etc. (ver por ejemplo [13]–[16]).

Recordamos la definición usual de una función de agregación (en este trabajo nos centramos en el intervalo unitario).

Definition 1: Un mapeo $f : [0,1]^n \to [0,1]$ es llamada función de agregación si satisface $f(0,\ldots,0) = 0$, $f(1,\ldots,1) = 1$ y monotonía creciente, p.ej., si $x_i \leq y_i$ para todo $i \in \{1,\ldots,n\}$, entonces $f(x_1,\ldots,x_n) \leq f(y_1,\ldots,y_n)$. *Observación:* en la literatura reciente, la propiedad de monotonía de las funciones de agregación ha sido extendida o generalizada (ver por ejemplo [17], [18]).

En este trabajo, solo nos centraremos en las funciones de agregación con comportamiento promediado, también llamadas *medias*. Estas funciones cumplen que $\min(x_1,\ldots,x_n) \leq f(x_1,\ldots,x_n) \leq \max(x_1,\ldots,x_n)$ para todo $(x_1,\ldots,x_n) \in [0,1]^n$. Observar también que, debido a la monotonía, el comportamiento promedio es equivalente al comportamiento idempotente, es decir, $f(x,\ldots,x) = x$ para todo $x \in [0,1]$.

Ejemplos bien conocidos de funciones de agregación promediadas son las siguientes:

- La media aritmética $AM(x_1,\ldots,x_n) = \frac{1}{n}(x_1 + \ldots + x_n)$;
- La mediana

$$MED(x_1,\ldots,x_n) = \begin{cases} \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right) & \text{si } n = 2k \\ x_{(k)} & \text{si } n = 2k-1 \end{cases}$$

  con $k \in \mathbb{N}^+$ y $x_{(k)}$ siendo k el elemento más largo (más pequeño) de $x_1,\ldots,x_n$;
- La media geométrica $GM(x_1,\ldots,x_n) = \sqrt[n]{x_1 x_2 \cdots x_n}$;
- La media armónica $HM(x_1,\ldots,x_n) = \frac{n}{\frac{1}{x_1}+\cdots+\frac{1}{x_n}}$.

Cabe decir que cuando se aplican las funciones de fusión de información en un problema del mundo real, se deben incorporar algunas fuentes de información (en adición a las propias entradas). Esto se puede hacer de manera sencilla aplicando funciones de agregación ponderadas, dado que los pesos nos permiten modelar la importancia de cada atributo o criterio a ser fusionado.

Definition 2: Un vector $\mathbf{W} = (w_1,\ldots,w_n)$ es un vector de ponderación si $w_i \in [0,1]$ y $\sum_{i=1}^n w_i = 1$.

A partir de la definición del vector de ponderación podemos definir fácilmente la media aritmética ponderada como $AM_{\mathbf{w}}(x_1,\ldots,x_n) = \sum_{i=1}^n w_i x_i$.Observar que cuando incorporamos el vector de ponderación se pierde la simetría de la media aritmética.

Otra familia importante de funciones de agregación ponderadas son los operadores OWA. En este caso, los pesos no son asociados con una entrada particular (criterio) sino con su magnitud. Por lo tanto, los operadores OWA son funciones de agregación simétricas. Dado un vector de pesos $\mathbf{W}$, el operador OWA es definido como $OWA_{\mathbf{w}}(x_1,\ldots,x_n) = \sum_{i=1}^n w_i x_{(i)}$donde $x_{(i)}$ es el elemento i más largo de $x_1,\ldots,x_n$.

II-A. De funciones de agregación ponderadas a basadas en medidas

Como ya hemos mencionado, los vectores de pesos son utilizados para dar importancia a cada entrada individual. Sin embargo, cuando trabajamos con problemas complejos, cada entrada (criterio, atributo o fuente de información) no esta totalmente aislada del resto de entradas. Esto quiere decir que claramente podemos tener interacciones positivas o negativas entre entradas. Un herramienta adecuada para modelar esta interacción son las medidas difusas (no aditivas) [19].

Definition 3: Sea $\mathcal{N} = \{1,\ldots,n\}$. Una medida difusa discreta es una función $m : 2^{\mathcal{N}} \to [0,1]$ que satisface las condiciones de frontera $m(\emptyset) = 0$, $m(\mathcal{N}) = 1$ y de monotonía, p.ej. $m(A) \leq m(B)$ siempre que $A \subset B$ para todo $A, B \subseteq \mathcal{N}$.

Para definir una medida difusa es necesario establecer sus $2^n - 2$ componentes. Esto puede ser una tarea compleja cuando n es bastante grande, por lo tanto, en la literatura se pueden encontrar varias simplificaciones. Una de las más conocidas son las λ medidas de Sugeno [9].

Definition 4: Sea $\lambda \in (-1,\infty)$. Podemos decir que $m : 2^{\mathcal{N}} \to [0,1]$ es una λ medida de Sugeno si, para todo $A, B \subseteq \mathcal{N}$ con $A \cap B = \emptyset$, entonces $m(A \cup B) = m(A) + m(B) + \lambda m(A)m(B)$.

Un paso más allá en la modelación del conocimiento en la fusión de información son las integrales difusas, como la integral de Choquet o la integral de Sugeno, que estan basadas en medidas difusas.

Definition 5: Dada una medida difusa $m : 2^{\mathcal{N}} \to [0,1]$, la integral discreta de Choquet es dada por

$$C_m(x_1,\ldots,x_n) = \sum_{i=1}^{n} (x_{\sigma(i)} - x_{\sigma(i-1)}) m(\{\sigma(i),\ldots,\sigma(n)\})$$

donde $\sigma : \mathcal{N} \to \mathcal{N}$ es una permutación tal que $x_{\sigma(1)} \leq \ldots \leq x_{\sigma(n)}$.

Observar que si m es aditiva, p. ej. para todo $A, B \subset \mathcal{N}$, $A \cap B = \emptyset$ entonces $m(A \cup B) = m(A) + m(B)$, la integral de Choquet es la media aritmética ponderada. Si m es simétrica, p. ej. para cualquier $A, B \subseteq \mathcal{N}$, $m(A) = m(B)$ donde $|A| =$

$|B|$, entonces la integral de Choquet es un operador OWA. Finalmente, si m es simétrica y aditiva, entonces la integral de Choquet es la media aritmética.

Definition 6: Dada una medida difusa $m : 2^{\mathcal{N}} \rightarrow [0,1]$, la integral discreta de Sugeno es dada por

$$S_m(x_1,\ldots,x_n) = \max_{i=1}^{n} \min\{x_{\sigma(i)}, m(\{\sigma(i),\ldots,\sigma(n)\})\}$$

donde $\sigma : \mathcal{N} \rightarrow \mathcal{N}$ es una permutación tal que $x_{\sigma(1)} \leq \ldots \leq x_{\sigma(n)}$.

III. Ensembles y el problema de las clases no balanceadas

En esta sección introduciremos los ensembles y el problema de las clases no balanceadas en el cual nos centramos en este trabajo.

III-A. Ensembles: construcción y agregación

Los ensembles están compuestos por un número de clasificadores y su objetivo principal es mejorar el rendimiento de los clasificadores individuales. No hay mejora posible cuando se combina el mismo clasificador y por tanto, se requieren formas de construir diferentes clasificadores a partir de los datos originales. Una vez estos clasificadores han sido entrenados, los nuevos ejemplos son clasificados sometiéndolos a todos los clasificadores y combinando sus salidas obteniendo la clase final para los ejemplos no etiquetados. La fase de la combinación también es conocida como fusión o agregación [1]. Este trabajo se centra en este aspecto de los ensembles de clasificadores, aunque primero tenemos que presentar como crear los diferentes clasificadores antes de pasar por su combinación.

Existen varios métodos para construir los ensembles de clasificadores, los cuales se centran principalmente en crear clasificadores base con un buen balance entre precisión y diversidad. Entre ellos, los algoritmos de aprendizaje mas utilizados son AdaBoost [3] y Bagging [2]. En ambos casos, los clasificadores son entrenados estratégicamente con el objetivo de conseguir la diversidad requerida.

Breiman [2] propuso el algoritmo Bagging, un método simple pero efectivo para construir ensembles. En este método, cada clasificador es entrenado con un subconjunto diferente de los datos originales. Por lo tanto, se utiliza un nuevo conjunto de datos para construir cada clasificador muestreando aleatoriamente (con reemplazo) ejemplos del conjunto original.

A la hora de clasificar nuevos ejemplos, se tienen en cuenta las salidas de todos los clasificadores. Normalmente, para obtener la clase final se utiliza la mayoría o la mayoría ponderada (donde se considera la confianza devuelta por el clasificador). La salida se obtiene utilizando la siguiente fórmula:

$$H(x) = I\left(\frac{1}{T}\sum_{t=1}^{T} h_t(x) > \theta\right) \quad (1)$$

donde $h_t \in [0,1]$ son los clasificador inducidos, I es la función indicador (devuelve 1 si la condición es satisfecha y 0 en otro caso) y θ es el umbral para decidir la clase (normalmente $\theta = 0{,}5$). Asumiendo que cada clasificador devuelve la confianza de su predicción (que se puede interpretar como la probabilidad de que el ejemplo pertenezca a la clase 1), esta formula simplemente refleja que esas predicciones son promediadas obteniendo la probabilidad media de que el ejemplo pertenezca a la clase 1 (es por ello que normalmente se considera $\theta = 0{,}5$). Esto es comúnmente conocido como la mayoría ponderada, mientras que el voto mayoritario se considera cuando $h_t \in \{0,1\}$.

Se ve claramente que la fórmula de promediado puede ser cambiada por cualquier otra fórmula presentada en la sección anterior. De hecho este es nuestro interés en este articulo. Tratamos de mejorar esta combinación teniendo en cuenta las interacciones de los clasificadores. Para ello, compararemos el uso de diferentes funciones de agregación: las no ponderadas (medias aritmética, geométrica y armónica), las ponderadas (media aritmética ponderada y los OWA) y los basados en medidas (Choquet y Sugeno). Obviamente, las últimas agregaciones son las que tienen en cuenta las interacciones, mientras que las no ponderadas directamente combinan las diferentes probabilidades y las ponderadas utilizan un único peso por cada clasificador. En el caso de las agregaciones ponderadas y las basadas en medidas, el establecer sus parámetros es clave. Este asunto se explica en la sección IV, donde proponemos una nueva forma de construir las medidas difusas a partir del rendimiento de los clasificadores.

III-B. El problema de las clases no balanceadas

Siempre que el numero de ejemplos de cada clase no sean parecidos, el conjunto de datos va a estar desequilibrado. Centrándonos en los problemas de dos clases, el problema es que una clase esta infrarrepresentada aunque desde el punto de vista del aprendizaje, esta suele tener mayor importancia [20]. En este contexto, los algoritmos de aprendizaje estándares tienden a posicionar sus clasificaciones hacia la clase mayoritaria, dado a su diseño orientado a precisión.

Existen diferentes métodos lidiar con el problema de las clases no balanceadas. En este trabajo no hemos centrado en las soluciones basadas en ensembles [10], y más concretamente, en el método conocido como UnderBagging [11] debido a su buen comportamiento. UnderBagging funciona introduciendo un re-muestreo aleatorio en cada iteración para crear un conjunto de datos (bolsa) balanceado. Por lo tanto, una bolsa es creada eliminando aleatoriamente ejemplos de la clase mayoritaria hasta tener un conjunto balanceado.

Otra cuestión importante cuando se trabaja con distribuciones de clases no balanceadas, es cómo medir la calidad del clasificador. En este escenario la precisión ya no es una medida válida ya que no es capaz de reflejar la precisión de ambas clases. Por esta razón, normalmente se construye la matriz de confusión de la cual se pueden obtener diferentes medidas. Entre ellas, el Ratio de Verdaderos Positivos ($\mathrm{TP}_{rate} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$) y el Ratio de Verdaderos Negativos ($\mathrm{TN}_{rate} = \frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}}$) permiten medir la calidad de la clasificación para clase por separado. Sin embargo, para ser capaces de comparar dos métodos, normalmente se prefiere evaluar el rendimiento de

ambas clases al mismo tiempo. Para ello, se puede utilizar la media geométrica (GM) [21], la cual se muestra en (2).

$$\mathrm{GM} = \sqrt{\mathrm{TP}_{rate} \cdot \mathrm{TN}_{rate}}. \quad (2)$$

IV. Estableciendo los pesos y las medidas difusas en base al rendimiento de los clasificadores

En esta sección presentamos el método seguido para establecer los parámetros de las funciones de agregación. Primero explicaremos como se calculan los pesos para la media aritmética ponderada y los OWA. Después presentaremos nuestra propuesta para el cálculo de la medida difusa basándonos en el rendimiento de los clasificadores.

IV-A. Pesos para las agregaciones ponderadas

Primero explicaremos cómo calcular los vectores de pesos para los operadores OWA. Recordar que para estas funciones, los pesos modelan la importancia de la magnitud de cada entrada. En este trabajo hemos seguido la metodología descrita en [22], donde los pesos son obtenidos por un cuantificador difuso $Q : [0,1] \to [0,1]$ utilizando la fórmula $w_i = Q\left(\frac{i}{n}\right) - Q\left(\frac{i-1}{n}\right)$ para todo $i \in \{1,\ldots,n\}$. Hemos utilizado tres cuantificadores lineales diferentes: Q_{alh} con $a = 0, b = 0{,}5$, Q_{amap} con $a = 0{,}5, b = 1{,}0$ y Q_{mot} con $a = 0{,}3, b = 0{,}8$, que son calculados por:

$$Q(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x > b \end{cases}$$

En el caso de la media aritmética ponderada, se incorpora conocimiento del problema que estamos tratando de resolver en el vector de pesos. Para ello, establecemos cada peso de cada clasificador utilizando el rendimiento normalizado obtenido por cada clasificador sobre el conjunto de entrenamiento. Por ejemplo, si $Per_1, \ldots, Per_n$ son los rendimientos de cada uno de los n clasificadores, entonces $w_i = \frac{Per_i}{\sum_{j=1}^{n} Per_j}$ para todo $i \in \{1,\ldots,n\}$. Dado que nos hemos centrado en conjuntos de datos no balanceados, utilizamos la GM como medida de rendimiento para modelar mejor la calidad de cada clasificador. Observar que en este caso, solo se tiene en cuenta la calidad individual de cada clasificador, y no se utiliza información sobre como interaccionan entre ellos.

IV-B. Propuesta para la construcción de la medida difusa

Para las integrales de Choquet y Sugeno necesitamos construir una medida difusa $m : 2^{\mathcal{N}} \to [0,1]$ con $\mathcal{N} = \{1,\ldots,n\}$, siendo n el número de clasificadores considerados. Recordamos que nuestro objetivo con esta medida es modelar las interacciones entre los clasificadores. Queremos reflejar el hecho de que un clasificador puede llegar a complementarse mejor con otro y, por lo tanto, llevarnos a una mejor solución. O, igualmente, que cuando se añade un nuevo clasificador, el desempeño del nuevo conjunto de clasificadores no incrementa como se esperaba. Dado que podemos estimar fácilmente como interaccionan los clasificadores evaluando cada posible subconjunto de clasificadores, proponemos utilizar su rendimiento para construir la medida difusa.

El proceso para establecer los valores de la medida difusa consiste en dos pasos. Primero, empezamos construyendo una medida difusa uniforme $m_U : 2^{\mathcal{N}} \to [0,1]$ la cual se calcula como $m_U(A) = \frac{|A|}{n}$ para todo $A \subseteq \mathcal{N}$. Esta medida será la base para le medida objetivo. Luego, en el segundo paso consideraremos el rendimiento de cada clasificador individual así como el rendimiento de cada posible combinación de clasificadores, llamado Per_A, para todo $A \subseteq \mathcal{N}$. Ahora para cada subconjunto con la misma cardinalidad, calculamos la media aritmética de sus rendimientos. Por ejemplos, para todo subconjunto de clasificadores con cardinalidad $i \in \{1,\ldots,n\}$, calculamos $MeanPer_i$. Finalmente, el valor de la medida difusa para cada $A \subseteq \mathcal{N}$ será dado por:

$$m(A) = m_U(A) + \frac{\tanh(100 \cdot (Per_A - MeanPer_{|A|}))}{2n} \quad (3)$$

donde $\tanh : (-\infty, +\infty) \to (-1,1)$ es la función tangente hiporbólica dada por $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. *Nota:* Como los rendimientos son números reales en el intervalo unitario, multiplicamos estos por 100 para considerar el rendimiento como porcentaje.

Observar que el mapeo m dado en la Ecuación 3 satisface $m(\emptyset) = 0$ y $m(\mathcal{N}) = 1$. Con respecto a la monotonía, es suficiente con ver que para todo $i = 2,\ldots,n-1$ y $A \subset \mathcal{N}$ con $|A| = i$, tenemos que $m(A) \in [(2i-1)/n, (2i+1)/n]$ y mantiene la monotonía. Analizando esta expresión, los rendimientos de los clasificadores que son mejores que el rendimiento medio en el mismo nivel serán incrementados y aquellos que son peores serán decrementados con respecto a la medida uniforme. Al igual que en el caso de la media ponderada, utilizamos como medida de rendimiento la GM.

Para comparar nuestra propuesta con métodos anteriores de medidas difusas y para mostrar las ventajas considerando todos los parámetros, utilizaremos las medidas difusas lambda [6], [7]. En este caso, solo tenemos que establecer el primer nivel de la medida difusa, dado que λ se obtiene de los valores de este nivel y posteriormente se puede derivar el resto de la medida difusa. El primer nivel de la medida difusa se obtiene de la misma forma que el método propuesto para reducir las diferencias entre su forma de trabajo.

V. Estudio experimental

En esta sección, primero explicaremos el marco experimental considerado para este estudio. Posteriormente, se prestan y se discuten los resultados experimentales.

V-A. Marco Experimental

Hemos considerado los treinta y tres conjuntos de datos mas desequilibrados (mayor ratio de no balanceo) del repositorio de datos de KEEL [12]. Hemos utilizado la validación cruzada con 5 particiones para obtener los resultados de cada método para cada conjunto de datos. Este esquema se repite 5 veces con diferentes semillas. Como medida de rendimiento utilizamos la GM (aunque se obtienen conclusiones similares cuando se considera la AUC). Además, hemos considerado el uso de test estadísticos no paramétricos [23] para analizar

adecuadamente los resultados obtenidos. Se ha utilizado el test de Friedman de rangos alineados para comparar un grupo de métodos con el objetivo de detectar diferencias significativas. En el caso de diferencias significativas, se aplica el test *post-hoc* de Holm para encontrar los algoritmos que rechazan la hipótesis nula de equivalencia frente al método de control seleccionado. Por otro lado, consideramos el test de Wilcoxon para estudiar si existen diferencias significativas entre dos métodos.

Siguiendo el análisis experimental desarrollado en trabajos anteriores, hemos considerado el árbol de decisión C4.5 [24] como clasificador base para nuestro ensemble (con pruning, nivel de confianza 0.25, mínimo número de ítems por hoja de 2 y confianza estimada mediante Laplace). Respeco al ensemble UnderBagging, hemos establecido el número de clasificadores a 10 para hacer posible la estimación de los valores de la medida difusa.

Para estudiar la ventaja de tener en cuenta la interacción entre clasificadores, compararemos el uso de diferentes funciones de agregación para la combinación de los clasificadores del ensemble. Realizaremos comparaciones intra- e inter-familiares con el objetivo de encontrar la mejor agregación. Todos los métodos considerados en el estudio se presentan en la Tabla I.

Tabla I
MÉTODOS CONSIDERADOS EN LA COMPARACIÓN

Family	Abb.	Description
Unweighted	AM	Media Aritmética
	MED	Mediana
	GM	Media Geométrica
	HM	Media Armónica
Weighted	AM_{gm}	Media aritmetica ponderada utilizando GM para el cálculo de los pesos
	Q_{alh}	OWA usando Q_{alh} como función para el cálculo de los pesos
	Q_{amap}	OWA usando Q_{amap} como función para el cálculo de los pesos
	Q_{mot}	OWA usando Q_{mot} como función para el cálculo de los pesos
Choquet	C_{gm}	Choquet usando la métrica GM para el calculo de los pesos
	$C_{\lambda_{gm}}$	Choquet usando medidas λ y GM como métrica de rendimiento
Sugeno	S_{gm}	Sugeno usando la métrica GM para el calculo de los pesos
	$S_{\lambda_{gm}}$	Sugeno usando medidas λ y GM como métrica de rendimiento

VI. Resultados y discusión

La Tabla II muestra los resultados obtenidos por cada método empleado (Tabla I) en términos de GM. Para cada conjunto de datos, se marca en **negrita** el mejor resultado. La última fila corresponde al rendimiento medio sobre todos los conjuntos de datos.

Atendiendo a la Tabla II, se puede observar que la función de agregación que obtiene mejor resultado medio es C_{gm}. Sin embargo, S_{gm} que usa la misma medida difusa no es capaz de conseguir el mismo rendimiento y obtiene una GM media menor que la media aritmética estándar. De todas formas, de estos números podemos concluir que las diferencias entre los diferentes métodos no son grandes en términos absolutos. No obstante, tenemos que estudiar estos resultados con el análisis estadístico adecuado.

Dado el número de métodos considerados en la comparación, hemos llevado a cabo un análisis estadístico jerárquico de los resultados, primero realizando comparación intra-familiares y posteriormente comparando los mejores métodos

Tabla II
GM MEDIA OBTENIDA POR CADA MÉTODO EN CADA CONJUNTO DE DATOS

	Unweighted				Weighted				Choquet		Sugeno	
Dataset	AM	MED	GM	HM	WAM_{gm}	Q_{alh}	Qamap	Q_{mot}	Cgm	$C\lambda_{gm}$	Sgm	$S\lambda_{gm}$
Glass04vs5	**0.9939**	**0.9939**	**0.9939**	**0.9939**	**0.9939**	**0.9939**	**0.9939**	**0.9939**	**0.9939**	**0.9939**	**0.9939**	**0.9939**
Ecoli0346vs5	0.8822	0.8784	0.8423	0.7260	0.8845	0.8225	0.8446	0.8855	0.8881	**0.8890**	0.8875	0.8884
Ecoli0347vs56	0.8659	0.8667	0.8060	0.7105	0.8666	0.8503	0.8126	0.8686	0.8737	0.8684	**0.8750**	0.8669
Yeast05679vs4	0.8112	0.8047	0.7837	0.6898	0.8088	0.7715	0.7695	0.8094	**0.8161**	0.8053	0.8070	0.8060
Ecoli067vs5	0.8733	0.8726	0.8410	0.7239	0.8738	**0.8800**	0.8245	0.8731	0.8740	0.8738	0.8733	0.8718
Vowel0	0.9600	0.9604	0.9402	0.9196	0.9600	**0.9681**	0.9438	0.9589	0.9614	0.9619	0.9627	0.9624
Glass016vs2	0.6566	0.6490	0.5494	0.3449	0.6605	0.5436	0.5438	0.6555	**0.6611**	0.6381	0.6528	0.6511
Glass2	0.7140	0.7180	0.5692	0.3465	0.7143	0.5253	0.5555	0.7087	**0.7287**	0.7161	0.7225	0.7205
Ecoli0147vs2356	0.8357	0.8365	0.7860	0.7002	0.8288	0.8235	0.7851	0.8262	0.8304	0.8354	0.8399	**0.8417**
Led7digit02456789vs1	0.8114	0.8150	0.7970	0.7879	0.8114	**0.8413**	0.7907	0.8081	0.8108	0.8195	0.8210	0.8210
Glass06vs5	0.9204	0.9183	0.9040	0.8581	0.9204	**0.9399**	0.9097	0.9130	0.9204	0.9324	0.9204	0.9345
Ecoli01vs5	0.8726	0.8718	0.8185	0.7091	0.8710	0.8637	0.8452	0.8659	**0.8770**	0.8768	0.8755	0.8759
Glass0146vs2	0.7060	0.7011	0.5494	0.3607	0.7092	0.5493	0.5456	0.6890	**0.7227**	0.7207	0.7073	0.7070
Ecoli0147vs56	0.8669	0.8630	0.7729	0.6510	0.8678	0.8637	0.7983	0.8563	0.8678	**0.8748**	0.8674	0.8722
Cleveland0vs4	0.8212	0.8085	0.6955	0.5763	0.8089	0.7249	0.7248	0.8152	**0.8268**	0.7778	0.8132	0.7752
Ecoli0146vs5	0.8777	0.8778	0.8102	0.6963	0.8661	0.8549	0.8195	0.8803	**0.8893**	0.8776	0.8874	0.8755
Ecoli4	0.8947	0.8943	0.8313	0.7345	0.8965	0.8975	0.8442	0.8906	0.8974	0.8965	0.8960	**0.8989**
Yeast1vs7	0.7498	0.7405	0.6366	0.5368	0.7470	0.6392	0.6128	0.7358	**0.7517**	0.7426	0.7508	0.7424
Shuttlec0vsc4	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
Glass4	**0.9171**	0.9040	0.8094	0.7062	0.9073	0.8731	0.8518	0.9052	0.9078	0.8985	0.9088	0.9096
Pageblocks13vs4	0.9731	0.9736	0.9095	0.8371	0.9785	**0.9881**	0.9327	0.9663	0.9790	0.9818	0.9797	0.9818
Abalone918	0.7512	**0.7539**	0.6847	0.5909	0.7534	0.6469	0.6668	0.7413	0.7522	0.7470	0.7467	0.7475
Glass016vs5	0.9411	0.9411	0.9411	0.9167	0.9411	**0.9441**	0.9411	0.9411	0.9411	0.7535	0.9411	0.9423
Shuttlec2vsc4	**1.0000**	**1.0000**	0.9812	0.9017	**1.0000**	0.9883	0.9930	**1.0000**	**1.0000**	0.9883	**1.0000**	**1.0000**
Yeast1458vs7	0.6132	0.6028	0.5034	0.4346	0.6171	0.4232	0.4265	0.5869	**0.6262**	0.6145	0.5932	0.6082
Glass5	0.9473	0.9473	0.9452	0.8880	0.9473	**0.9680**	0.9473	0.9473	0.9473	0.9554	0.9473	0.9590
Yeast2vs8	0.7337	0.7301	0.6914	0.5726	0.7320	0.7382	0.6817	0.7388	**0.7398**	0.7255	0.7253	0.7303
Yeast4	0.8387	0.8345	0.8098	0.7818	0.8386	0.8262	0.8123	0.8338	**0.8455**	0.8427	0.8386	0.8398
Yeast1289vs7	0.7168	0.7071	0.6037	0.4960	0.7237	0.6260	0.5801	0.7073	**0.7275**	0.7153	0.7196	0.7104
Yeast5	0.9495	0.9504	0.9270	0.9044	0.9496	**0.9628**	0.9373	0.9526	0.9513	0.9513	0.9479	0.9487
Ecoli0137vs26	0.7745	0.7724	0.6510	0.4892	**0.7762**	0.7561	0.7055	0.7504	0.7637	0.7371	0.7161	0.7093
Yeast6	0.8618	**0.8624**	0.8155	0.7471	0.8623	0.8521	0.8286	0.8566	0.8610	0.8602	0.8619	0.8622
Abalone19	0.6950	0.6902	0.6225	0.5792	0.6985	0.5827	0.6114	0.6850	**0.7076**	0.7063	0.6899	0.7007
Average	0.8432	0.8406	0.7825	0.6943	0.8429	0.8039	0.7842	0.8378	**0.8467**	0.8357	0.8415	0.8411

de cada familia. La Figura 1 representa este estudio. Cuando la comparación involucra solo dos métodos (familias Choquet y Sugeno), aplicamos el test de Wilcoxon, cuyos rangos se presentan cerca de cada método (a mayor el rango, mejor es el método). En otro caso, se utiliza el test de rangos alineados de Friedman para la comparación (a menor rango, mejor es el método). Además, remarcamos en **negrita** los rangos cuando existen diferencias significativas con respecto al método ganador (con $\alpha = 0{,}05$). Para ello, en el caso del test de Friedman, utilizamos el post-hoc test de Holm.

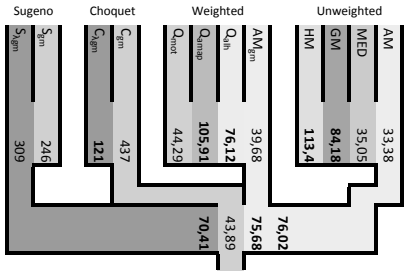

Figura 1. Estudio estadístico jerárquico comparando las funciones de fusión en cada familia y la mejor de cada familia

Entre las agregaciones no ponderadas, la media aritmética (AM) es la que mejor resultados obtiene seguida muy de cerca por la mediana (MED), mientras que la GM y la HM no proporcionan resultados competitivos. Respecto a las agregaciones ponderadas, AM_{gm} supera estadísticamente a todas excepto a un OWA (OWA_{mot}), lo cual muestra la importancia de considerar la calidad de cada clasificador para realizar la agregación de los clasificadores en los ensembles basados en Bagging (con respecto a considerar únicamente la magnitud del voto para establecer el peso). Finalmente, las

agregaciones de Choquet y Sugeno presentan dos escenarios diferentes. Mientras que en Choquet, la medida propuesta se comporta mejor, en la familia Sugeno la medida lambda consigue un mejor rendimiento. No obstante, al enfrentar el mejor método de cada familia, el claro ganador es C_{gm}. La agregación Choquet con la medida propuesta, la cual sí tiene en cuenta la interacción entre los clasificadores, es capaz de superar al resto. De hecho, Choquet presenta un mejor resultado que la otra alternativa de integral difusa.

Finalmente, para completar nuestro estudio experimental, hemos comparado directamente el mejor método de cada familia frente a la AM, ya que es la agregación más comúnmente considerada para combinar clasificadores. Estas comparaciones se han realizado utilizando el test de Wilcoxon y se presentan en la Tabla III. Por cada comparación se presentan los rangos a favor de cada método, el p-valor y si la hipótesis nula de equivalencia es rechazada o no.

Tabla III
Test de Wilcoxon comparando la AM contra la mejor función de fusión de cada familia

Comparación	R^+	R^-	Hipótesis ($\alpha = 0{,}05$)	p-value
AM vs. AM_{gm}	214	311	No rechazada	0.2348
AM vs. C_{gm}	89	451	**Rechazada para C_{gm}**	0.0006
AM vs. $S_{\lambda_{gm}}$	252	303	No rechazada	0.6106

En esta tabla se puede observar que C_{gm} es la única agregación capaz de superar estadísticamente a la AM. Es más, este resultado es interesante ya que muestran que el uso de la AM puede ser mejorado con la función de agregación adecuada.

VII. Conclusiones

En este trabajo hemos presentado una nueva forma para aprender una medida difusa a partir del rendimiento de los clasificadores con el objetivo de mejorar la fase de agregación de los ensembles. Hemos realizado un estudio experimental completo comparando nuestra propuesta con diferentes familias de funciones de agregación y también con el uso de las medidas difusas lambda consideradas en la literatura. En el marco experimental considerado, los experimentos han presentado que la medida propuesta junto a la integral de Choquet mejora significativamente el uso de la comúnmente utilizada media aritmética.

En el futuro nuestro objetivo es estudiar cómo abordar la mayor desventaja del uso de las medidas difusas, el hecho de que hay que estimar 2^N parámetros (siendo N el número de entradas a agregar). Además, nuestro propósito es mejorar el marco experimental añadiendo mas conjuntos de datos y otros escenarios diferentes al del problema de las clases no balanceados. También compararemos el método propuesto con otros métodos para la agregación de clasificadores que no están directamente basados en el uso de funciones de agregación.

Agradecimientos

Este trabajo ha sido apoyado en parte por el Ministerio Español de Ciencia y Tecnología bajo el Proyecto TIN2016-77356-P (AEI/FEDER, UE).

Referencias

[1] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.
[2] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996.
[3] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.
[4] J.-L. Marichal, "On choquet and sugeno integrals as agregation functions," in *Fuzzy Measures and Integrals. Theory and Applications*, T. M. M. Grabisch and M. Sugeno, Eds. Physica-Verlag, 2000, pp. 247–272.
[5] ——, "Aggregation of interacting criteria by means of the discrete choquet integral," in *Aggregation Operators. New Trends and Applications*, T. Calvo, G. Mayor, and R. Mesiar, Eds. Physica-Verlag, 2002, pp. 224–244.
[6] S.-B. Cho and J. H. Kim, "Multiple network fusion using fuzzy logic," *IEEE Transactions on Neural Networks*, vol. 6, no. 2, pp. 497–501, 1995.
[7] L. I. Kuncheva, ""fuzzy"versus "nonfuzzyïn combining classifiers designed by boosting," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 6, pp. 729–741, 2003.
[8] G. Choquet, "Theory of capacities," *Ann. Inst. Fourier*, vol. 5, pp. 1953–1954, 1953.
[9] M. Sugeno, "Theory of fuzzy integrals and its applications," Ph.D. dissertation, Tokyo Institute of Technology, 1974.
[10] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, vol. 42, no. 4, pp. 463–484, 2012.
[11] R. Barandela, R. M. Valdovinos, and J. S. Sánchez, "New applications of ensembles of classifiers," *Pattern Analysis & Applications*, vol. 6, pp. 245–256, 2003.
[12] J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17:2-3, pp. 255–287, 2011.
[13] T. Calvo, G. Mayor, and R. Mesiar, *Aggregation Operators. New Trends and Applications*. Physica-Verlag, 2002.
[14] G. Beliakov, A. Pradera, and T. Calvo, *Aggregation Functions: A Guide for Practitioners*. Springer, 2007.
[15] M. Grabisch, J.-L. Marichal, R. Mesiar, and E. Pap, *Aggregation Functions*. Cambridge University Press, 2009.
[16] G. Beliakov, H. Bustince, and A. Pradera, *A Practical Guide to Averaging Functions*, 2nd ed. Springer, 2015.
[17] H. Bustince, J. Fernandez, A. Kolesárová, and R. Mesiar, "Directional monotonicity of fusion functions," *European Journal of Operational Research*, vol. 244, pp. 300–308, 2015.
[18] D. Paternain, M. Campión, H. Bustince, I. Perfilieva, and R. Mesiar, "Internal fusion functions," *IEEE Transactions on Fuzzy Systems*, InPress.
[19] Y. Narukawa and V. Torra, "Fuzzy measure and probability distributions: distorted probabilities," *IEEE Transactions on Fuzzy Systems*, vol. 13, pp. 617–629, 2005.
[20] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.
[21] R. Barandela, J. S. Sánchez, V. García, and E. Rangel, "Strategies for learning in class imbalance problems," *Pattern Recognition*, vol. 36, no. 3, pp. 849–851, 2003.
[22] R. Yager, "Quantifier guided aggregation using owa operators," *International Journal of Intelligent Systems*, vol. 11, pp. 49–73, 1998.
[23] S. García, A. Fernández, J. Luengo, and F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power," *Information Sciences*, vol. 180, pp. 2044–2064, 2010.
[24] J. R. Quinlan, *C4.5: Programs for Machine Learning*, 1st ed. San Mateo-California: Morgan Kaufmann Publishers, 1993.