UNA REVISIÓN DE LOS MÉTODOS DE AGREGACIÓN DE CLASIFICADORES.

Esteban Alfaro Cortés Matías Gámez Martínez Noelia García Rubio
Área de Estadística. Departamento de Economía y Empresa.
Universidad de Castilla-La Mancha.
Plaza de la Universidad, s/n. 02071 Albacete.
e-mail: {Esteban.Alfaro, Matías.Gámez, Noelia.Garcia}@uclm.es

RESUMEN.

En este artículo se pretende plantear las ventajas que presenta la combinación de distintos clasificadores individuales para conseguir una mayor precisión de forma conjunta. Se estudian las razones fundamentales por las que se puede explicar la superioridad de estos métodos de agregación, que son básicamente tres, una razón estadística, una de computación y otra de representación. Se recogen algunas de las posibles clasificaciones de los métodos de combinación. Se analiza el sistema *Bagging* como método de agregación que crea sus clasificadores base entrenando un clasificador sobre distintas muestras bootstrap del conjunto de entrenamiento. También se estudia el método *Boosting* basado en la construcción de sucesivos clasificadores sobre el conjunto de entrenamiento que se va modificando en función de los errores cometidos por el clasificador anterior. Finalmente se recogen algunas de las comparaciones que se han realizado en diversos estudios.

PALABRAS CLAVE: Combinación de clasificadores, bagging, boosting.

1. INTRODUCCIÓN.

Un individuo preocupado por su estado de salud decide ir al médico para saber qué enfermedad padece. El doctor después de preguntarle qué es lo que le pasa y realizarle las pruebas que considere necesarias le dará un diagnóstico. Si el médico es capaz de realizar el diagnóstico es porque el paciente presenta alguno, sino todos, de los síntomas que se asocian a una enfermedad. Para ello es necesario que previamente se hayan estudiado gran cantidad de pacientes con esa misma enfermedad y con otras similares y llegar a determinar qué síntomas son los caracterizan exclusivamente a una enfermedad en concreto y no a las similares.

Esta labor tan importante es lo que se conoce como reconocimiento de patrones o problema de clasificación. En clasificación se parte de un conjunto de observaciones de clase conocida, que se conoce como conjunto de entrenamiento, y se debe ser capaz de determinar aquellas características que diferencian una clase del resto. De esta manera se podrá asignar correctamente una de las clases posibles cuando se presente una nueva observación de clase desconocida.

Por tanto, inicialmente se tiene un conjunto de *N* observaciones que se representa por $T_n = \{(X_1,Y_1), (X_2,Y_2), ..,(X_n,Y_n)\}$ donde cada X_i es un vector *p*-dimensional que recoge los valores que presenta la observación *i*-ésima en cada una de las *p* características o atributos utilizados, es decir, $X_i = \{X_{i1}, X_{i2}, ..., X_{ip}\}$e Y es la etiqueta de la clase a la que pertenece esa observación y toma valores en {1, 2, ..., *k*}. A partir de este conjunto de entrenamiento, se construye un clasificador que de forma general será una función de las *p* variables observables, $f(X_i)$, que se utiliza para predecir el valor de Y de tal forma que el error cometido sea lo más pequeño posible.

El tipo de aprendizaje expuesto, donde se conoce a priori cuáles son las posibles clases, se conoce como *aprendizaje supervisado*. Por el contrario cuando no se conocen las clases y el objetivo es precisamente establecer grupos con rasgos diferenciados dentro del conjunto de observaciones, se habla de *aprendizaje no supervisado*.

En clasificación también hay que distinguir entre *métodos paramétricos y no paramétricos*. En el primer caso se parte de una estructura funcional conocida para $f(X_i)$ y el objetivo es determinar únicamente los parámetros de esa función que minimicen el error. En el segundo tipo, los métodos no paramétricos, hay que determinar también la estructura funcional para conseguir un mejor ajuste y por tanto un menor error. Para una descripción más detallada acerca del

problema de clasificación ver Alfaro, Gámez y García (2002).

En caso de que el enfermo anterior no confíe plenamente en el diagnóstico proporcionado por ese médico, es posible que decida pedir una segunda opinión a otro doctor para ver si coincide y quedar así más tranquilo. Si pudiese quizás consultase no a dos médicos sino a un grupo de ellos, un comité, para asegurarse de que todos diagnostican lo mismo. El problema puede agravarse en caso de que no coincidan los diagnósticos de los distintos doctores, ¿qué debe hacer el pobre paciente?.

En realidad tiene varias posibilidades, las más sencilla sería elegir aquél diagnóstico que se haya repetido más veces entre los facultativos consultados, lo que se llamaría el diagnóstico mayoritario. Pero quizás no todos los médicos le inspiren la misma confianza y por tanto para él tenga un mayor valor la opinión de unos que la de otros, a la hora de elegir el diagnóstico definitivo.

Esto por supuesto sólo es un ejemplo, pero es aplicable también a los problemas de clasificación, para aumentar la precisión se puede recurrir a un comité de clasificadores, es decir, a la agregación de las predicciones de varios clasificadores. Cuando se habla de agregación se está haciendo referencia a la combinación de varios clasificadores. El clasificador resultante de la combinación de varios clasificadores, lo llamaremos clasificador *combinado.* Existen varias alternativas, entre ellas construir distintos clasificadores sobre el conjunto de datos disponible y combinarlos después mediante votación simple o funciones lineales. Otra posibilidad, quizás más sofisticada, consiste en aplicar un mismo método de clasificación sobre versiones modificadas del conjunto de entrenamiento. Algunas de estas técnicas son relativamente novedosas y han sido muy estudiadas en los últimos años, entre ellas se pueden citar los métodos *bagging* y *boosting*.

2. ¿POR QUÉ UTILIZAR COMBINACIONES DE CLASIFICADORES?

La idea de partida es utilizar un conjunto de clasificadores para obtener una mayor precisión de la que cada uno de éstos logra de manera individual. Cada método de clasificación se basa en conceptos o procedimientos de estimación diferentes, además puesto que todos los métodos de

clasificación tienen algún punto fuerte o ventaja sobre la regla por defecto[1] es lógico intentar aunar las mejores propiedades de cada uno de ellos combinándolos de alguna manera. Para ello en ocasiones se trabaja con la clase predicha por los clasificadores individuales mientras que en otras, lo que se utiliza son las probabilidades condicionadas asignadas a cada clase por los distintos clasificadores.

Combinar la salida de varios clasificadores es útil únicamente si hay desacuerdo entre ellos. Obviamente combinando varios clasificadores idénticos no se obtiene ningún beneficio. Hansen y Salomon (1990) probaron que si la tasa de error media para una observación es menor del cincuenta por ciento y los clasificadores utilizados en el comité son independientes en la producción de sus errores, el error esperado para una observación puede ser reducido a cero cuando el número de clasificadores combinado se acerca a infinito.

Por su parte, Krogh y Vedelsby (1995) probaron más tarde que el error conjunto puede dividirse en un término que mide el error de generalización medio de cada clasificador individual y un término que recoge el desacuerdo entre los clasificadores. Lo que ellos demostraron formalmente fue que la combinación ideal consiste en clasificadores con alta precisión que estén el mayor número de veces posible en desacuerdo.

Volviendo al planteamiento de Hansen y Salomon, incluso el método más sencillo de combinación, el voto mayoritario y bajo el supuesto de que los clasificadores son independientes entre sí, se puede comprobar para un problema dicotómico como la precisión del conjunto es superior a la de los clasificadores individuales, siempre que estos cometan un error inferior al cincuenta por ciento. En concreto puede verse como la probabilidad de que de todos los clasificadores individuales considerados fallen más de la mitad. El gráfico 1 muestra como evoluciona el error del conjunto en función del número de clasificadores utilizados en la combinación para diferentes valores de la probabilidad de error de los clasificadores individuales (ε). Se considera únicamente las combinaciones con un número impar de clasificadores, como suele hacerse en la práctica para evitar empates. A medida que aumenta el número de clasificadores utilizados aumenta la precisión del conjunto para un valor dado de la precisión

[1] La regla por defecto consiste en asignar a todas las observaciones la clase mayoritaria en el conjunto de entrenamiento sin tener en cuenta ningún atributo o característica de las que presenta esa observación.

individual.

Por su parte el gráfico 2 recoge cómo varía el error del conjunto en función de la precisión de los clasificadores base, entre 0 y 0,5, para distintos tamaños del conjunto . Como se podía esperar cuanto menor es la precisión de los clasificadores básicos menor es también la del conjunto para un mismo tamaño del conjunto. Cualquiera que sea el número de clasificadores utilizados, cuando su precisión alcanza el valor 0,5 no se obtiene ningún beneficio a partir de esta combinación.

A pesar de la sencillez y claridad de esta explicación, está basada en supuestos muy restrictivos en la práctica como es la independencia entre los clasificadores básicos. Además el planteamiento para el cálculo de la probabilidad de error del conjunto sólo es válido para éste método de combinación. Por todo ello, deben existir otras razones más consistentes que expliquen la superioridad del conjunto sobre los clasificadores individuales. En Dietterich (2000) se plantean las tres razones siguientes.

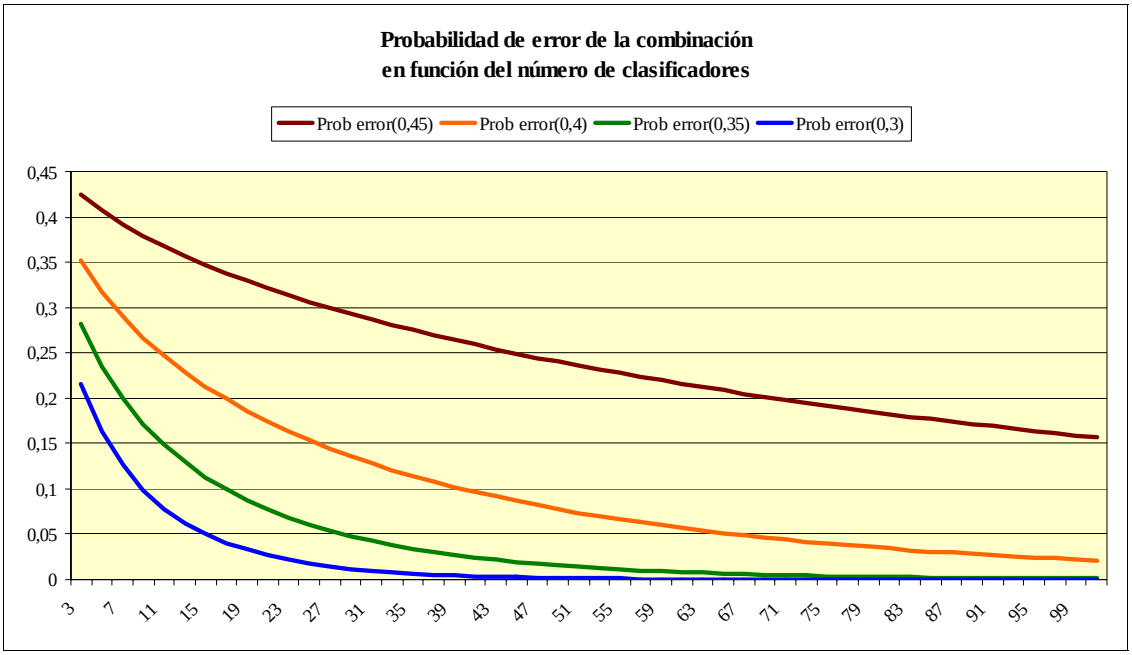

Gráfico 1. Probabilidad de error de la combinación mediante voto mayoritario en función del número de clasificadores utilizados en la misma.

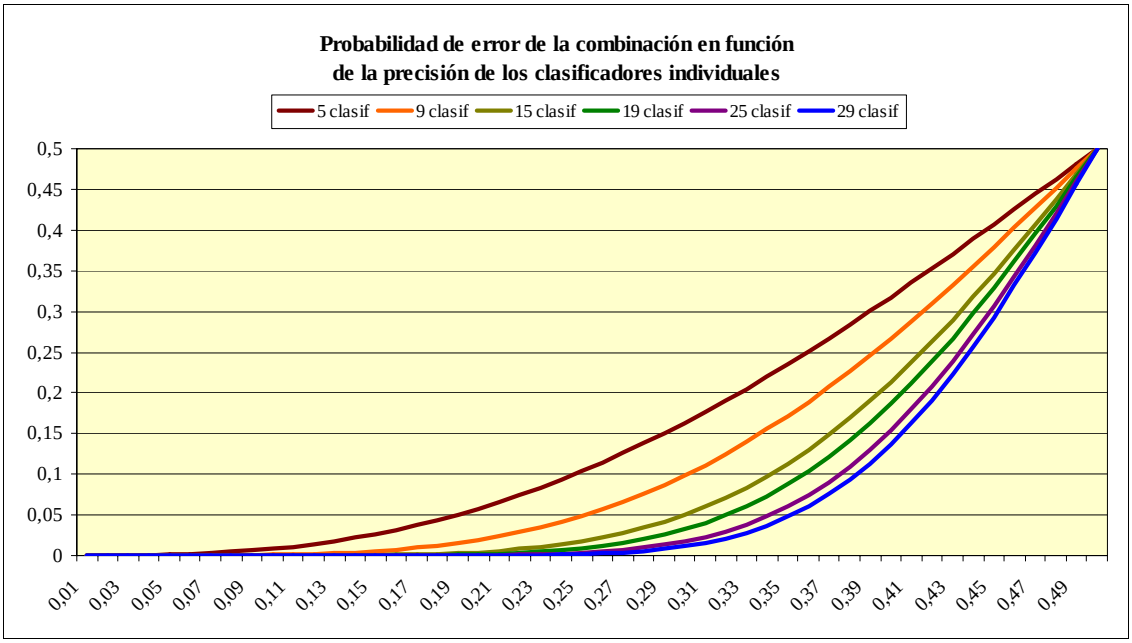

Gráfico 2. Probabilidad de error de la combinación mediante voto mayoritario en función de la precisión de los clasificadores individuales.

La primera razón Dietterich la denomina razón estadística. Un sistema de aprendizaje puede entenderse como la búsqueda dentro de un determinado espacio de hipótesis de aquella que sea la más adecuada. En la práctica es habitual encontrarse con conjuntos de datos cuyo tamaño es demasiado pequeño en comparación con el tamaño del espacio de hipótesis. En estos casos el método de clasificación puede encontrarse con varias hipótesis distintas que presentan una precisión similar en el conjunto de datos disponible. Si se combinan las distintas hipótesis, se pueden promediar los resultados y reducir el riesgo de elegir un clasificador que obtenga una menor precisión a la hora de generalizar.

En segundo lugar considera una razón de computación. Viene causada por la propia naturaleza de determinados sistemas de clasificación que realizan algún tipo de búsqueda local por lo que pueden quedar atrapados en un óptimo local. En este tipo de clasificadores se encuentran entre otros las redes neuronales que utilizan el método del gradiente descendente para minimizar una función de coste sobre las observaciones de entrenamiento y los árboles de clasificación que utilizan una regla de corte secuencial en el desarrollo del árbol. Incluso en casos donde el tamaño del conjunto de entrenamiento es suficiente como para que no exista el problema anterior, sigue siendo computacionalmente difícil para el sistema de aprendizaje encontrar la mejor hipótesis. En conclusión, la combinación de clasificadores obtenidos realizando la búsqueda local desde

puntos iniciales distintos logrará una mejor aproximación a la hipótesis o función objetivo buscada que cualquiera de los clasificadores individuales.

La tercera es la razón de representación. En muchos problemas de clasificación la verdadera función no se puede representar mediante ninguna de las hipótesis existentes en el espacio de hipótesis disponibles. Sin embargo, mediante combinaciones de hipótesis relativamente sencillas de ese espacio se puede llegar a una mejor aproximación a la función real mediante la expansión del espacio de funciones representables. Por ejemplo, utilizando una suma ponderada de las hipótesis simples. Esta cuestión es aplicable incluso a aquellos métodos de clasificación que teóricamente tienen capacidad para representar cualquier función. Por ejemplo, las redes neuronales o los árboles de clasificación que son métodos muy flexibles y con una gran capacidad de representación cuando el número de observaciones es ilimitado. Pero en la práctica lo habitual es enfrentarse a conjuntos de datos limitados por lo que incluso estos sistemas trabajan con un conjunto de hipótesis también limitado y detendrán la búsqueda cuando encuentren una hipótesis que se ajuste correctamente a los datos disponibles.

Estas son las tres razones fundamentales que señala Dietterich para justificar la superioridad de la combinación de clasificadores sobre los individuales. Ahora bien no se debe olvidar que también presentan algunas desventajas como son la pérdida de comprensibilidad (estructura más compleja y mayor tamaño), la mayor lentitud en la construcción de la combinación (se necesita una mayor capacidad de memoria) y también se tarda más a la hora de clasificar una nueva observación.

3. CLASIFICACIÓN DE LOS MÉTODOS DE COMBINACIÓN.

Debido a la importancia que recientemente se le ha dado a la agregación de métodos en el reconocimiento de patrones, en la última década se han propuesto varias clasificaciones de los métodos de combinación de los sistemas de aprendizaje. Aunque existen otras se han seleccionado sólo algunas de ellas.

En primer lugar, se puede citar la clasificación propuesta por Kittler (1998) que distingue según el punto de vista del análisis, básicamente dos escenarios de combinación. En el primero, todos los clasificadores utilizan la misma representación en los patrones de entrada u observaciones. Por ejemplo, un conjunto de clasificadores del vecino más próximo que utilizan

el mismo vector de variables pero distinto valor de *k* o en la distancia empleada para determinar el número de vecinos a considerar. O un conjunto de redes neuronales con la misma estructura pero con diferente vector de pesos obtenidos con diferentes estrategias de aprendizaje. De esta forma cada clasificador, para una observación determinada, se puede entender que las probabilidades a posteriori que obtiene para cada clase son una estimación de una misma función objetivo.

En el segundo caso, cada clasificador utiliza su propia representación de las observaciones de entrada. Es decir, las medidas extraídas de las observaciones son únicas para cada clasificador. Una importante aplicación de la agregación de clasificadores en este escenario es la integración de distintos tipos de características o atributos de las observaciones. En esta situación, ya no es posible considerar las probabilidades a posteriori calculadas como estimación de un mismo valor funcional, puesto que los sistemas de clasificación trabajan en diferentes espacios de medida.

En Dietterich (2000) se establecen los siguientes grupos de los métodos de combinación diferenciando entre aquellos que realizan una votación de tipo Bayesiano, los que modifican los ejemplos de entrenamiento, los que modifican las variables, los que modifican las clases posibles, y por último los que aleatorizan el sistema de aprendizaje.

En tercer lugar, Lam (2000) propone agrupar estos métodos según la arquitectura de la agregación distinguiendo entre si ésta se realiza en serie, en paralelo o de forma jerárquica. En Jain, Duin y Mao (2000) recoge la separación en función de si los clasificadores básicos son seleccionados o no por el algoritmo de combinación, diferenciando entre los métodos de combinación orientados a la selección y aquellos orientados a la combinación.

Por último, Masulli y Valentini (2002) Utilizan una clasificación parecida a la anterior en función de si el algoritmo de combinación actúa o no sobre los clasificadores básicos modificándolos, de esta forma distingue entre métodos generadores y no generadores de combinación. Los *métodos generadores* crean conjuntos de clasificadores básicos actuando sobre el propio sistema de clasificación o sobre el conjunto de datos de entrenamiento intentando activamente mejorar la diversidad y la precisión de los clasificadores básicos. Los *métodos no generadores* se limitan a combinar un conjunto dado de clasificadores básicos posiblemente bien diseñados, es decir, no generan nuevos clasificadores básicos, sino que intentan combinar de la

mejor forma posible los ya existentes.

En **los métodos no generadores** los clasificadores básicos se unen mediante un procedimiento de combinación que depende de su capacidad de adaptación a las observaciones de entrada y de las necesidades de la salida que proporcionan los sistemas de aprendizaje individuales. Es decir, el tipo de combinación depende del tipo de salida. Si sólo se dispone de la clase asignada o si las salidas continuas son difíciles de manejar, entonces se utiliza el voto mayoritario, ésta es la forma más sencilla de combinar varios clasificadores y consiste en asignar a una observación la clase que predicen la mayoría de los clasificadores. Es decir, cada nueva observación hay que presentarla ante cada clasificador individual para ver qué clase le asigna. Una vez que todos los clasificadores individuales han dado su predicción se asigna a esa nueva observación la clase que mayoritariamente se haya repetido. En caso de empate se asigna la clase con mayor probabilidad a priori y si ésta no se conoce, la clase mayoritaria en el conjunto de entrenamiento o en el peor de los casos se resuelve al azar.

El problema reside en que de esta forma todos los clasificadores del comité tienen la misma importancia y no se tiene en cuenta la mayor o menor precisión del mismo a la hora de la generalización. Además el número de clasificadores incluido será un factor crítico ya que demasiados clasificadores poco precisos pueden llevar a equivocar la decisión final y por tanto a disminuir la precisión del conjunto.

Este procedimiento puede mejorarse asignando un peso a cada clasificador individual de tal forma que se optimiza el comportamiento del clasificador combinado en el conjunto de entrenamiento. Si los clasificadores básicos proporcionan las probabilidades a posteriori de las clases se pueden agregar utilizando operadores sencillos como el mínimo, el máximo, la media, la mediana, el producto o la media ponderada.

Los métodos generadores intentan mejorar la precisión global de la combinación mediante la actuación directa sobre la precisión y la diversidad de los clasificadores base. Dentro de este grupo se pueden distinguir subgrupos en función de las estrategias que utilizan para conseguir mejorar los clasificadores básicos.

1. Métodos de selección de variables. En este caso se reduce el número de características utilizadas en los clasificadores básicos, de esta forma sirve para hacer frente al problema de la

dimensionalidad (Skurichina, 2001, p.156), que consiste en la escasez de datos en relación al número de variables que describen cada una de las observaciones. Estos métodos de subespacios de características actúan dividiendo el conjunto de atributos, utilizando cada subconjunto para entrenar un clasificador básico. El método más utilizado es el Método del Subespacio Aleatorio[2] propuesto por Ho (1998), que selecciona aleatoriamente un subconjunto de características donde se entrena el clasificador base. De esta forma se obtiene un subespacio aleatorio del espacio original de características y se construye el clasificador en ese subespacio. La agregación se realiza normalmente mediante voto ponderado por la precisión de los clasificadores individuales. En Skurichina (2001) se muestra que este método es efectivo para clasificadores cuya curva de aprendizaje es decreciente y que estén construidos en conjuntos de entrenamiento de tamaño pequeño y crítico.

2. Métodos de prueba y selección. Se basan en la idea de seleccionar los clasificadores básicos durante el proceso de creación de la combinación. Aunque existen otras alternativas, cabe destacar la selección hacia delante y hacia atrás a imitación de las estrategias que se siguen para la selección de variables en algunos sistemas como el análisis discriminante. Consiste en un proceso secuencial donde en cada paso sólo se incluye (o se extrae) un nuevo clasificador del comité si esto conlleva una reducción en el error.

3. Métodos aleatorios de agregación. Estos procedimientos aleatorizan el algoritmo de aprendizaje para generar combinaciones de sistemas de aprendizaje. Un ejemplo sería iniciar con valores aleatorios los pesos en el algoritmo de retropropagación, obteniendo clasificadores distintos para utilizarlos en la combinación. Algunos resultados experimentales muestran que aleatorizar los sistemas de aprendizaje utilizados para generar los clasificadores básicos de la combinación mejora el comportamiento de los clasificadores individuales no aleatorios, Dietterich (2000).

1. Métodos de remuestreo. Estas técnicas de remuestreo pueden utilizarse para extraer muestras distintas a partir del conjunto de datos original de tal forma que los clasificadores individuales entrenados en cada una de ellas sean utilizados posteriormente en la combinación. De entre la técnicas de remuestreo la que más se utiliza para este propósito es la técnica de

[2] Este método se conoce en inglés como Random Subspace Method (RSM).

bootstrap para obtener muestras sin reemplazamiento del mismo tamaño que el conjunto de datos original, de tal forma que se obtienen distintos conjuntos de entrenamiento. Estas técnicas son especialmente aconsejables para sistemas de clasificación inestables, es decir, métodos muy sensibles a pequeños cambios en el conjunto de entrenamiento, como pueden ser los árboles de clasificación y las redes neuronales. Dentro de éstos métodos de combinación que utilizan el remuestreo los más utilizados son los métodos *Bagging* y *Boosting* de los que se habla a continuación.

4. BAGGING.

Uno de los problemas más habituales a la hora de establecer un clasificador para un conjunto de datos es el tamaño limitado del conjunto de ejemplos de entrenamiento. Aunque este problema afecte especialmente a los métodos paramétricos supone un reto para cualquier clasificador. Cuanto más pequeño sea el conjunto de datos disponibles menos seguro se puede estar de que este conjunto represente fielmente a la población total. En general los clasificadores construidos en conjuntos pequeños pueden estar sesgados y presentarán una elevada varianza en la probabilidad de clasificación errónea. Se dice en este caso que el clasificador es inestable.

En muchos casos no se puede disponer de más observaciones y por tanto el conjunto de entrenamiento está limitado. Existen diversas técnicas que intentan obtener clasificadores más estables y actualmente este es uno de los campos de investigación abiertos en el ámbito de sistemas de clasificación.

Una posible solución es utilizar el sistema Bagging (**B**ootstrapping and **agg**regat**ing**). De igual forma que la estimación bootstrap de los parámetros de la distribución de los datos es más precisa y robusta que la estimación tradicional, se puede pensar en el uso de ésta técnica para conseguir, una vez combinado, un clasificador con mejores propiedades.

Bagging fue propuesto por Breiman en 1996 y se basa en los métodos de bootstrapping y de agregación. Tanto los métodos de bootstrapping como los de agregación presentan propiedades beneficiosas. Bootstrapping consiste en obtener muestras aleatorias con reemplazamiento de igual tamaño que el conjunto original. Partiendo del conjunto de entrenamiento $X= (X_1, X_2, ..., X_n)$, mediante la extracción aleatoria con reemplazamiento con el mismo número de elementos que el conjunto original de *n* elementos, se obtienen B muestras bootstrap $X^b= (X_1^b, X_2^b, ..., X_n^b)$

donde b=1, 2,.., B. En algunas de estas muestras se habrá eliminado o al menos reducido la presencia de observaciones ruidosas, por lo que el clasificador construido en ese conjunto presentará un mejor comportamiento que el clasificador construido en el conjunto original. Así pues Bagging puede ser útil para construir un mejor clasificador cuando el conjunto de entrenamiento presente observaciones ruidosas.

El clasificador *combinado*, obtiene frecuentemente mejores resultados que los clasificadores individuales utilizados para construir el clasificador final. Esto puede entenderse si se considera que al combinar los clasificadores individuales se están combinando las ventajas de cada uno de ellos en el clasificador final.

En concreto el método Bagging se aplica del siguiente modo:

1. Repetir para b=1, 2, ...,B

a) Realizar una réplica bootstrap X^b del conjunto de entrenamiento X.

b) Construir un clasificador sencillo $C_b(x)$ en X^b (con frontera de decisión igual a $C_b(x)=0$)

2. Combinar los clasificadores básicos $C_b(x)$, b=1,2,...B usando el voto mayoritario (la clase predicha más frecuente) en la regla de decisión final

$$\beta(x)=\arg\max_{y\in\{-1,1\}} \sum_b \delta_{sgn}(C_b(x),y) \quad donde\ \delta(i,j)=\begin{cases}1 & i=j\\ 0 & i\neq j\end{cases} \qquad \textbf{(1)}$$

Existen otras funciones de combinación basadas en probabilidades a posteriori: mínimo, máximo, media, mediana y producto.

Otra posibilidad es promediar los coeficientes de los clasificadores básicos, para lo que se utiliza la función promedio, presenta las siguientes ventajas, en primer lugar no es necesario retener todos los resultados de la clasificación, sólo los coeficientes. En segundo lugar obtiene un único clasificador con el mismo número de parámetros que los clasificadores básicos.

Cuando se extrae una réplica bootstrap del conjunto de entrenamiento inicial X= (X_1, X_2, ..., X_n), la probabilidad de que la observación i-ésima X_i (i=1,2,...,n) sea incluida *m* veces (m=1,2,..,n) en esa muestra bootstrap X^b vendrá dada por la distribución binomial B (*n*, *1/n*)

donde *1/n* es la probabilidad que tiene X_i de ser seleccionada en cada extracción y *n* es el número de extracciones con reposición que se efectúan, en concreto la probabilidad de que X_i sea extraida m veces será:

$$P(m)=\binom{n}{m}\left(\frac{1}{n}\right)^{m}\left(1-\frac{1}{n}\right)^{n-m} \quad \textbf{(2)}$$

Cuando $1/n<0'1$, es decir para conjuntos de más de diez observaciones se puede aproximar la distribución binomial a través de la Poisson, $P(\lambda)$, donde $\lambda=n(1/n)=1$, y en consecuencia la probabilidad de que de que X_i sea extraída *m* veces es:

$$P(m)=\frac{e^{-1}}{m!} \quad \textbf{(3)}$$

De este modo cada observación tiene una probabilidad aproximada de *1/e* de no ser incluida (sustituyendo en la ecuación anterior m=0) en una réplica bootstrap. Por lo tanto, a la larga podemos esperar que, en término medio, aproximadamente el 37% de las observaciones se quedarán fuera de una muestra bootstrap. De esta forma las posibles observaciones ruidosas del conjunto de entrenamiento no aparecerán en algunas de esas muestras. En ese caso, el clasificador construido bajo esas condiciones obtendrá mejor tasa de error aparente que el construido en el conjunto de entrenamiento original con observaciones ruidosas y lógicamente debe tener una mayor influencia en la decisión final que otras versiones bootstrap. De esta forma se explica que el clasificador obtenido por bagging, consiga mejores resultados que los clasificadores individuales.

En realidad los clasificadores construidos en las muestras bootstrap de los conjuntos de entrenamiento unas veces obtienen un clasificador mejor que el original y en otras uno peor que éste. El clasificador que sea superior al original tendrá una mayor probabilidad a posteriori y, por ello, dominará en la decisión final. Por lo tanto, la combinación de las versiones bootstrap del clasificador nos permite obtener un clasificador mejor que el original.

Las diferentes versiones del bagging han sido estudiadas por varios investigadores, aunque quizás con más frecuencia en el ámbito de la regresión que desde el punto de vista de la

clasificación. En Breiman (1996) se muestra que bagging puede reducir el error tanto de regresión como de clasificación en los árboles de decisión. Aunque según este autor bagging es beneficioso para clasificadores inestables únicamente. Mientras que para procedimientos estables puede incluso llegar a perjudicar el comportamiento del clasificador. Por ejemplo, para el método de clasificación del vecino más próximo.

Skurichina (2001) defiende que la estabilidad de los clasificadores lineales depende del tamaño del conjunto de entrenamiento. Por lo tanto, no se podría decir que bagging no es útil para un determinado clasificador de forma general, sino que dependerá de la aplicación concreta a la que se enfrente.

5. BOOSTING.

Como ya se ha dicho, dado un conjunto de datos un sistema de aprendizaje genera un clasificador capaz de predecir la clase de una nueva observación. La mayor o menor precisión de ese clasificador dependerá de la calidad del método utilizado y de la dificultad que presente la aplicación concreta. Siempre que el clasificador obtenido supere a la regla por defecto querrá decir que ha sido capaz de encontrar alguna estructura en los datos para conseguir esta ventaja. Boosting es un método que aumenta la precisión de un clasificador sacando provecho de su ventaja. De tal forma que utiliza el método de clasificación como una subrutina para producir un clasificador que consigue una alta precisión en el conjunto de entrenamiento.

Boosting aplica el sistema de clasificación varias veces sobre el conjunto de entrenamiento, pero cada vez dirige la atención del aprendizaje a diferentes ejemplos del mismo. Una vez que el proceso ha terminado, los clasificadores básicos obtenidos se combinan en un único clasificador final que será muy preciso en el conjunto de entrenamiento. El clasificador final normalmente logra también una precisión elevada en el conjunto de test, según han demostrado diversos autores tanto teórica como empíricamente.

Aunque existen diversas versiones de algoritmos boosting la más extendida es la que proporcionan Freund y Schapire (1996) que se conoce como AdaBoost. Para simplificar se puede suponer que sólo existen dos clases sin pérdida de generalidad. Se parte del conjunto de entrenamiento $T_n = \{(X_1,Y_1), (X_2,Y_2), ..,(X_n,Y_n)\}$ donde Y toma en este caso los valores$\{-1, 1\}$. Se asigna a cada observación X_i el peso $w_b(i)$ que inicialmente se iguala a 1/n y que

posteriormente se irá actualizando en cada iteración. Se construye un clasificador básico a partir de ese conjunto de entrenamiento que se representa por $C_b(X_i)$ y se aplica a cada uno de los ejemplos de entrenamiento. El error de ese clasificador se representa por ε_b y se calcula como

$$\varepsilon_b = \sum_{i=1}^{N} w_b(i)\,\xi_b(i) \quad donde\ \xi_b(i) = \begin{cases} 0 & C_b(x_i) = y_i \\ 1 & C_b(x_i) \neq y_i \end{cases} \qquad \textbf{(4)}$$

A partir del error del clasificador en la iteración b-ésima se calcula la constante α_b que se utiliza para la actualización de los pesos. En concreto estos autores hacen $\alpha_b = ½ \ln(1 - \varepsilon_b / \varepsilon_b)$ y el nuevo peso para la iteración b+1 será

$$w_{b+1}(i) = w_b(i) \exp(\alpha_b \xi_b(i)) \qquad \textbf{(5)}$$

posteriormente se normalizan los pesos calculados para que la suma de todos ellos sea uno. Según ellos $\varepsilon_b = 0{,}5 - \gamma_b$, donde γ_b representa la ventaja que presenta el clasificador básico de la iteración b sobre la regla por defecto en el peor de los casos en que las dos clases tengan la misma probabilidad a priori, 0'5.

Se puede ver el siguiente ejemplo de cómo se actualizan los pesos en función del error cometido para la segunda iteración, b=2. Si ε_b =0,499, entonces α_b = 0,002 y el nuevo peso para la segunda iteración será $w_2(i)=1/n \exp(0{,}002\ \xi_b(i))$. Luego si la observación i-ésima está clasificada incorrectamente el peso $w_2(i)=1/n \cdot 1{,}002$, mientras que si ha sido clasificada correctamente su peso se verá disminuido al normalizar. Si se considera ahora que ε_b =0,001, entonces α_b = 3,453 y el peso de una observación mal clasificada será $w_2(i)=1/n \cdot 31{,}6$, mientras que los pesos de las observaciones bien clasificadas será reducido en la normalización .

Como se puede ver se aumenta la ponderación de las observaciones mal clasificadas y se disminuye las de las clasificadas correctamente, forzando así al clasificador básico construido en la siguiente iteración a centrarse en aquellos casos que han resultado más difíciles. Además las diferencias en la actualización son mayores cuando el error cometido por el clasificador básico es pequeño, porque si el clasificador consigue una precisión elevada se le da más importancia a los pocos fallos cometidos.

Este proceso se repite en todas las iteraciones desde b=1, 2, 3, ..,B. Para acabar se construye el clasificador final como combinación lineal de los clasificadores básicos ponderados por la constante α_b correspondiente. En concreto será:

$$C(x)=sign\left(\sum_{i=1}^{B}\alpha_b C_b(x)\right) \qquad \textbf{(6)}$$

Freund y Schapire demostraron que al aumentar el número de iteraciones B, el error de entrenamiento del clasificador combinado Adaboost tiende a cero un ritmo exponencial. Estos autores probaron también que el error de generalización o real (ε_R) del clasificador final $C_B(x)$ tiene un límite superior que depende del error de entrenamiento o aparente (ε_A), del tamaño del conjunto de entrenamiento(n), del coeficiente de dimensionalidad de Vapnik-Chervonenkis (VC) del espacio paramétrico de los clasificadores básicos (*d*) y del número de iteraciones *B* utilizadas en boosting (número de clasificadores básicos combinados)

$$\hat{\varepsilon}_R=\hat{\varepsilon}_A+\hat{\theta}\sqrt{\frac{Bd}{n}} \qquad \textbf{(7)}$$

lo que se interpreta como que se puede disminuir el error de generalización de $C_B(x)$ aumentando el tamaño del conjunto de entrenamiento. Pero además también se ve que el error de generalización aumentará cuando aumente el número de clasificadores básicos incluidos. Esto significa que el clasificador final estará sobreajustado[3]

Comparación de bagging y boosting.

Los dos métodos construyen los clasificadores básicos en versiones modificadas del conjunto de entrenamiento y los combinan después en la regla de clasificación final. Sin embargo, se diferencian en el modo de obtener los clasificadores básicos. En concreto las dos principales diferencias son:

1. Todas las observaciones se utilizan en cada paso del boosting (no se aplica bootstraping),

[3]Se dice que un clasificador está sobreajustado cuando está demasiado adaptado al conjunto de entrenamiento y por tanto pierde capacidad de generalización a la población total, es decir, será poco preciso ante observaciones nuevas. Para una descripción más detallada de este problema ver Alfaro, Gámez y García (2002).

al menos en la versión inicial del mismo en Freund y Schapire (1996).

2. El clasificador obtenido en cada paso del boosting depende de todos los anteriores, mientras que en Bagging son independientes.

6. ALGUNAS COMPARACIONES REALIZADAS.

A pesar del enorme interés que han suscitado los métodos de combinación de clasificadores en la última década y de los muchos trabajos realizados sobre este tema, existen pocas comparaciones entre los distintos métodos. En realidad es muy difícil llevar a cabo una comparación exhaustiva debido a que el abanico de posibilidades es muy amplio. Hay que seleccionar qué métodos de combinación se van a estudiar y qué clasificadores básicos se van a utilizar, así como el número de éstos, porque no es lo mismo comparar boosting y bagging aplicados sobre análisis discriminante lineal que sobre árboles de clasificación o redes neuronales.

Entre los estudios más interesantes, está el realizado por Kuncheva, Bezdek y Duin (2001) en el que se compara los métodos de voto mayoritario y los de máximo, mínimo, media y producto. Para ello utiliza dos problemas de la base de datos ELENA (ftp anónimo en ftp.dice.ucl.ac.be, directorio *pub/neural-nets/ELENA/databases*), los conjuntos de fonemas e imágines de satélite. El primero de ellos presenta 5404 observaciones con cinco características y dos clases distintas. En el segundo se trabaja con 6435 ejemplos y únicamente cuatro atributos y seis clases. Como clasificador básico se utiliza el discriminante cuadrático (DC.) entrenado para cada par de variables por lo que en el caso del conjunto de los fonemas se obtienen diez clasificadores distintos y en el de imágenes seis. Se utilizan cuatro tamaños para el conjunto de entrenamiento 100, 200, 1000 y 2000 en ambos conjuntos y el resto de observaciones se utilizan como conjunto de prueba. La tabla 1 muestra los resultados obtenidos.

Tamaño de entrenamiento	Imágenes de satélite				Fonemas			
	100	200	1000	2000	100	200	1000	2000
D.C.	77,5	79,73	80,67	80,62	74,29	75,45	75,52	75,17
Voto May	80,89	81,27	82,13	82,23	75,51	75,96	76,38	76,08
Máximo	81,44	82,19	82,91	82,84	75,42	75,40	75,80	75,47
Mínimo	82,51	83,53	84,42	84,57	75,42	75,40	75,80	75,47
Media	82,63	82,85	83,81	83,88	75,82	75,81	76,37	75,91
Producto	83,02	83,55	84,44	84,48	75,77	75,81	76,34	75,88

Tabla 1. Fuente: Kuncheva, Bezdek y Duin (2001)

Puede observarse como las combinaciones superan en precisión a los mejores clasificadores individuales casi en todos los casos. Pero no existe una supremacía general de ninguno de los métodos de combinación considerados, en el problema de las imágenes domina el producto en tres de los cuatro casos, mientras en el caso de los fonemas es el voto mayoritario el vencedor en tres ocasiones. Los autores muestran su sorpresa ante el discreto comportamiento de la media considerado frecuentemente como el favorito entre éstos métodos sencillos.

Una segunda comparación relevante se presenta en Breiman (1996) en este caso se comparan los métodos de Bagging y Boosting (Adaboosting) en cinco conjuntos disponibles en la base datos de la Universidad de California, Irvine (UCI repository). Como clasificadores base se utilizan árboles de clasificación construidos mediante el sistema CART. Se comparan los porcentajes de error en el conjunto de prueba de estos métodos para un número de iteraciones determinado a priori (B =50) y para el caso en que se detiene el procedimiento cuando se alcanza un error de entrenamiento nulo.

Conjuntos de datos	Boosting		Bagging	
	B=50	error=0	B=50	error=0
heart	1,1	5,3	2,8	3,0
breast cancer	3,2	4,9	3,7	4,1
ionosphere	6,4	9,1	7,9	9,2
diabetes	26,6	28,6	23,9	24,7
glass	22,0	28,1	23,2	25,0

Tabla 2. Fuente: Breiman (1996)

Estos resultados confirman que boosting reduce el error de entrenamiento con mayor rapidez que bagging pero que esa reducción no se produce tan rápidamente en el error de prueba. Además puede verse como boosting logra una mayor reducción del error para un mismo número de

iteraciones salvo para el problema de diabetes.

En tercer lugar cabe destacar la evaluación empírica de bagging y boosting que realizan Maclin y Opitz (1997) utilizando 23 conjuntos de la base de datos de la universidad de California mencionada anteriormente. Este estudio tiene la virtud de comparar estos dos métodos utilizando como clasificadores base árboles de clasificación (CART) y redes neuronales. En el caso de las redes neuronales también compara una combinación de varios clasificadores obtenidos a partir de valores iniciales aleatorios de los pesos. La tasa de aprendizaje se fija en 0,15 y el término momento en 0,9. El número de nodos ocultos se elige en función del número de unidades de entrada y salida de cada problema y el número de iteraciones se basa en el número de ejemplos y en la estructura de la red.

Conjuntos de datos		Red Neuronal				CART		
		RN	Pesos aleat	bagging	boosting	CART	bagging	boosting
1	breast cancer	3,3	3,4	3,3	3,9	5	3,3	3,1
2	credit-a	14,8	14	14,1	16,2	14,9	12,1	12,6
3	credit-g	28,3	24,4	24,3	26,4	29,6	22,8	22,9
4	diabetes	23,6	22,8	23,2	22,8	28,3	21,9	22,3
5	glass	38,5	35,5	33,7	33,2	30,9	28,4	30,5
6	heart-cleveland	18,2	17,3	16,7	19,1	24,3	18,1	17,4
7	hepatitis	19,9	19,6	18,1	18,1	21,6	16,5	13,8
8	house-votes-84	5	4,9	4,3	5,1	3,5	3,6	4,4
9	hypo	6,4	6,2	6,2	6,2	0,5	0,4	0,4
10	ionosphere	10,1	8	7,6	8	8,1	6	6
11	iris	4,3	4	4,3	3,3	6	4,6	5,6
12	kr-vs-kp	2,3	0,9	0,9	0,4	0,6	0,5	0,3
13	labor	5,3	4,2	4,9	5	15,1	13,3	13,2
14	letter	18	12,8	12,5	4,6	14	10,6	6,7
15	promoters-936	5	4,8	4,5	4,7	12,8	9,5	6,3
16	ribosome-blind	9,5	8,5	8,4	8,5	11,2	9,3	9,1
17	satellite	12,9	11,1	11	10,3	13,8	10,8	10,4
18	segmentation	6,7	5,6	5,3	3,7	3,7	2,8	2,3
19	sick	6	5,7	5,8	4,8	1,3	1	0,9
20	sonar	16,9	16,7	16,5	12,5	29	21,6	19,7
21	soybean	9	6,4	6,8	6,3	8	8	7,9
22	splice	4,7	4	3,9	4,3	5,9	5,7	6,3
23	vehicle	24,5	21,1	21,7	19,5	29,4	26,1	24,8

Tabla 3. Fuente Maclin y Opitz (1997)

Para cada uno de los 23 conjuntos utilizados se ha señalado la casilla del método con menor tasa de error en el conjunto de prueba, los métodos que resultan vencedores en más ocasiones son boosting aplicado tanto a redes neuronales como a árboles de clasificación y bagging aplicado a CART. Solamente en una ocasión resulta ganador uno de los sistemas individuales, luego se constata la superioridad de los métodos de combinación sobre los métodos de clasificación individuales. Ahora bien, la elección de un método de combinación no parece sencilla a la luz de estos resultados ya que ningún método mantiene una supremacía de forma general sobre el resto como se puede ver en el gráfico 3.

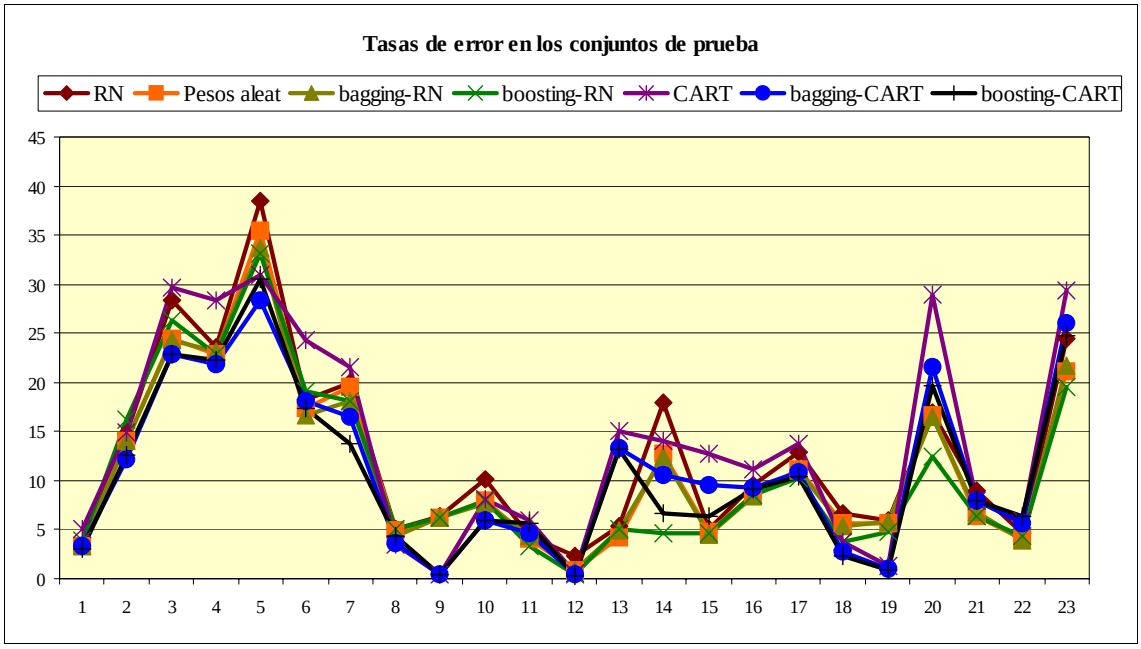

Gráfico 3. Tasas de error en los conjuntos de prueba.

7. CONCLUSIONES.

En este artículo se han estudiado los métodos de combinación de clasificadores, poniendo de manifiesto, tanto a nivel teórico como a nivel experimental, la ganancia de precisión que se consigue mediante algunas de estas técnicas. Se han establecido algunas clasificaciones de los métodos de agregación a partir de ciertas características que éstos presentan, como son el hecho de que los clasificadores combinados utilizan la misma representación o no para las observaciones de entrada, la arquitectura de la agregación, si selecciona los clasificadores básicos o simplemente los combina y, finalmente, si el método genera nuevos clasificadores individuales o se limita a combinarlos.

Se han analizado en particular dos técnicas de agregación, bagging y boosting. El sistema *Bagging* crea sus clasificadores base entrenando un sistema de clasificación sobre distintas muestras bootstrap del conjunto de entrenamiento. En cambio *Boosting* se basa en la construcción de sucesivos clasificadores sobre modificaciones del conjunto de entrenamiento realizadas en función de los errores cometidos por el clasificador anterior, concentrando su esfuerzo sobre los ejemplos más difíciles del conjunto de entrenamiento.

Además en las comparaciones empíricas se muestra la superioridad de los métodos de combinación sobre los métodos de clasificación individuales. Ahora bien, ningún método de combinación se muestra superior al resto de manera general. Existen muchas cuestiones importantes que no se han estudiado en este trabajo como pueden ser el efecto que tiene en la precisión del conjunto la no independencia entre los clasificadores combinados, la respuesta de los métodos de combinación en presencia de observaciones ruidosas y probar distintas combinaciones de clasificadores básicos.

REFERENCIAS.

ALFARO, E.; GÁMEZ, M. Y GARCÍA, N. (2002): *"Una Revisión de los métodos de clasificación aplicables a la Economía"*. Documento de Trabajo de la Facultad de CC. Económicas y Empresariales de Albacete.

BREIMAN, L. (1996): *"Bagging predictors"*. Machine Learning, Vol 24, 2, p.123-140.

BREIMAN, L. (1998). Arcing classifiers. The Annals of Statistics, Vol 26, 3, p. 801-849.

DIETTERICH, T.G. (2000): *"Ensemble methods in machine learning"*. En Multiple Classifier Systems, Cagliari, Italia.

FREUND, Y. Y SCHAPIRE, R.E. (1996): *"Experiments with a New Boosting Algorithm"*. En Proceedings of the Thirteenth International Conference on Machine Learning, p. 148-156, Morgan Kaufmann.

HANSEN, L. Y SALOMON, P. (1990): *"Neural Networks Ensembles"*. IEEE Trans. Pattern Analysis and Machine Intelligence, 12, p. 993-1001.

HO, T.K. (1998): *"The random subspace method for constructing decision forests"*. IEEE

Transactions on Pattern Analysis and Machine Intelligence, 20(8), p. 832--844.

JAIN, A.K.; DUIN, R.P.W.Y MAO J.C. (2000): *"Statistical pattern recognition: a review"*. IEEE Trans. Pattern Analysis and Machine Intelligence, 22(1), p. 4-37.

KITTLER, J.; HATEF, M.; DUIN, R.P.W. Y MATAS, J. (1998): *"On Combining Classifiers"*. IEEE Trans. Pattern Analysis and Machine Intell., 20, p. 226-238.

KROGH, A. Y VEDELSBY, J. (1995): *"Neural Networks Ensembles, Cross Validation and Active Learning."* En Toretzky, D.; Tesauro, G. y Leen, T.(ed): Advances in Neural Information Processing Systems, vol. 7, p. 107-115. MIT Press, Cambridge, MA.

KUNCHEVA, L.I.; BEZDEK, J.C. Y DUIN, R.P.W. (2001): *"Decision templates for multiple classifier fusion: and experimental comparison"* Pattern Recognition, 34, p.299-314.

LAM, L. (2000): *"Classifier combinations: implementations and theoretical issues"*. En Kittler, J. y Roli, F. (ed): Multiple Classifier Systems, vol 1857 de Lecture Notes in Computer Science, p. 78-86, Cagliari, Italia. Springer.

MACLIN, R. Y OPITZ, D. (1997): *"An empirical evaluation of bagging and boosting"*. En Proceedings of the Fourteenth National Conference on Artificial Intelligence, p. 546--551 Cambridge, MA. AAAI Press/MIT Press.

SKURICHINA, M. (2000): *"Stabilizing Weak Classifiers"*. Tesis Doctoral, Delft University of Technolog , Delft, Holanda.

VALENTINI, G. Y MASULLI, F. (2002): *"Ensembles of learning machines"*. En Marinaro, M. y Tagliaferri, R. (ed) Neural Nets WIRN Vietri, Series Lecture Notes in Computer Sciences, Springer-Verlag, Heidelberg, Alemania.