MÉTODOS DE COMBINACIÓN DE CLASIFICADORES UTILIZANDO MEDIDAS DE DIVERSIDAD

COLECTIVO DE AUTORES

Edición: Liset Ravelo Romero
Corrección: Estrella Pardo Rodríguez
Diagramación: Roberto Suárez Yera

Editorial Feijóo, 2014

ISBN: 978-959-250-897-2

Editorial Samuel Feijóo, Universidad Central “Marta Abreu” de Las Villas, Carretera a Camajuaní, km 5 ½, Santa Clara, Villa Clara, Cuba. CP 54830

Tabla de contenidos

1. Introducción

Con independencia del área de conocimiento en la que se esté trabajando, es frecuente la necesidad de identificar las características que permiten diferenciar a dos o más grupos de sujetos. Usualmente se requiere además poder clasificar nuevos casos como pertenecientes a uno u otro grupo: ¿se beneficiará este paciente del tratamiento o no?, ¿devolverá este cliente el crédito o no?, ¿se adaptará este candidato al puesto de trabajo o no? Estos constituyen apenas algunos ejemplos.

A falta de otra información, cualquier profesional se limita a utilizar su propia experiencia o la de otros, o su intuición, para anticipar el comportamiento de un sujeto: el paciente se beneficiará del tratamiento, el cliente devolverá el crédito o el candidato se adaptará a su puesto de trabajo en la medida en que se parezcan a los pacientes, clientes o candidatos que se benefician del tratamiento, que devuelven el crédito o que se adaptan a su puesto de trabajo. Pero a medida que los problemas se hacen más complejos y las consecuencias de una mala decisión más graves, las impresiones subjetivas basadas en la propia intuición o experiencia deben ser sustituidas por argumentos más consistentes. Tales argumentos son con frecuencia modelos matemáticos o de inteligencia artificial llamados clasificadores.

Numerosos son los clasificadores reportados en la literatura: análisis discriminante, árboles de decisión, redes neuronales, bayesianas y regresión logística o multinomial por sólo mencionar algunos.

Habitualmente los problemas de clasificación se resuelven con éxito utilizando un solo clasificador, por ejemplo: (Chavez, 2008).

En la actualidad existen campos multidisciplinarios como la Bioinformática, con problemas grandes y muy complejos, que no se logran solucionar de forma satisfactoria con el uso de un solo clasificador. Con frecuencia ocurre que la precisión, exactitud o la característica que se desee medir, de un único clasificador no satisface los requerimientos del problema. Esa es la razón principal que ha conllevado al auge en el uso conjunto de sistemas compuestos por varios clasificadores para tratar de alcanzar resultados

superiores a los de un clasificador. El propósito del presente trabajo es realizar un estudio de las diferentes técnicas existentes en esta área para que sirva de punto de partida para estudiantes e investigadores que deseen conocer más sobre el tema.

Los sistemas multiclasificadores logran satisfacer muchas veces la necesidad de desarrollar clasificadores exactos, precisos y confiables para muchas aplicaciones prácticas. Se han reportado numerosos artículos que los han utilizado con éxito en la solución de problemas reales. (Bonet, 2008).

El desarrollo de nuevos métodos clasificadores, así como nuevas vías de combinación junto con el aumento exponencial de las referencias bibliográficas en este campo, apoyan sobradamente la idea de realizar estados del arte periódicos en este tema.

La idea inicial para mezclar conjuntos de clasificadores es muy sencilla. Ella parte del hecho de combinar respuestas que se complementen unas con otras. Los clasificadores que deben combinarse no son, contrario a lo que pudiera pensarse, los más precisos o los más exactos, sino los más diversos. Una muestra mal clasificada por uno o varios clasificadores puede estar correctamente clasificada por otro u otros.

Mezclar un grupo de clasificadores idénticos no producirá mejores resultados que uno solo de ellos. La idea es entonces combinar un grupo de clasificadores diferentes entre sí, para garantizar que al menos uno de ellos de la respuesta correcta cuando el resto falle. Por tal razón resulta sumamente importante estudiar la diversidad de los clasificadores bases a combinar.

Existen numerosas fórmulas reportadas en la literatura, conocidas como “medidas de diversidad”. Ninguna de ellas es superior a otra, luego comprenderlas para cuantificar la diversidad que existente en un conjunto de clasificadores es un aspecto imprescindible en la combinación de clasificadores.

El objetivo fundamental de la presente monografía es presentar un estado del arte actualizado y crítico en el tema de clasificación, que incluya una breve explicación acerca de los clasificadores más frecuentemente utilizados, así como de sus combinaciones. Para complementar se enuncian y explican las medidas de diversidad más utilizadas reportadas en la literatura.

El trabajo está dividido en varias sesiones. La primera de ellas recoge las ideas esenciales de un conjunto de clasificadores que se han utilizado ampliamente en la solución en problemas reales con éxito: algoritmos basados en casos, árboles de decisión con diferentes criterios de segmentación, redes bayesianas, máquinas de soporte vectorial, redes neuronales artificiales, el clásico análisis discriminante y la regresión logística. Le sigue un estudio profundo y un análisis crítico de los diversos modelos de combinación de clasificadores reportados en la literatura, entre los que pueden mencionarse: bagging, boosting, stacking y voto mayoritario entre otros.

A continuación se presentan y discuten críticamente las medidas de diversidad que aprueban o no la combinación de un conjunto de clasificadores bases. Se presenta una clasificación de las mismas en medidas por pares o grupales y cada una de ellas se explica ampliamente.

El último epígrafe se dedica a comentar las ideas esenciales del problema de la forma de combinación de las salidas de los clasificadores.

A modo de resumen se enuncian las conclusiones del trabajo y se presenta la lista de la bibliografía consultada que sirve de material auxiliar para aquellos estudiantes, profesores o investigadores en un sentido general que deseen profundizar en los temas tratados.

2. Métodos de clasificación

Clasificación es la acción o el efecto de ordenar o de disponer por clases. (*www.wikipedia.com*). Los métodos matemáticos de clasificación pertenecen al llamado "aprendizaje supervisado". Ellos están caracterizados fundamentalmente porque se conoce la información acerca de la clase a la que pertenece cada uno de los objetos. Cuando la variable de decisión, función o hipótesis a predecir es continua, a los algoritmos relacionados con los problemas supervisados se les conoce como métodos de regresión. Si por el contrario la variable de decisión, función o hipótesis es discreta, ellos se conocen como métodos de clasificación o simplemente clasificadores. Este trabajo se centra en estos últimos.

En un problema de clasificación se tienen un conjunto de objetos, elementos, instancias u observaciones divididos en clases o etiquetados. Dado un elemento del conjunto, un especialista le asigna una clase de acuerdo a los rasgos, características o variables que lo describen. Esta relación entre los descriptores y la clase puede estar dada por un conjunto de reglas. La mayoría de las veces este conjunto de reglas no se conoce y la única información que se tiene es el conjunto de ejemplos etiquetados, de forma tal que las etiquetas representan las clases.

De manera general, se puede decir que los métodos de clasificación son un mecanismo de aprendizaje, donde la tarea es tomar cada instancia y asignarla a una clase particular. Las clases entre las que puede elegir el procedimiento de clasificación se pueden describir de gran cantidad de formas. Su definición dependerá del problema en particular. Este tipo de método constituye una parte importante de muchas de las tareas de resolución de problemas.

La clasificación puede dividirse en tres procesos fundamentales: pre-procesamiento de los datos, selección del modelo de clasificación y, entrenamiento y prueba del clasificador (Bonet, 2008).

Entre los métodos de clasificación más usados están los algoritmos basados en casos, árboles de decisión, redes bayesianas, máquinas de soporte vectorial, redes neuronales artificiales, análisis discriminante y regresión logística, pero estos no son los únicos. A continuación se presenta una breve descripción década uno de ellos.

1.1 Algoritmos basados en casos

El razonamiento basado en casos se basa en el principio de usar experiencias viejas para resolver problemas nuevos. Muchos algoritmos usan este razonamiento para resolver los problemas y entre los más comunes están los de clasificación.

Aunque todos los métodos de clasificación se basan en casos, existe un conjunto que se conoce como algoritmos basados en casos, o también como métodos de aprendizaje perezoso. Estos algoritmos deben contar con una serie de ejemplos ya conocidos y cuando van a resolver un problema nuevo, lo hacen buscando la semejanza entre éste y los ejemplos almacenados, no necesitan crear reglas, ni árboles, ni ajustar parámetros. A cada ejemplo se le conoce como instancia y a la colección de ejemplos como base de casos.

Una nueva instancia se compara con el resto de la base de casos a través de una medida de similitud. La clase de la nueva instancia será la misma que la del caso que más cercano esté a la nueva instancia. A este proceso se le conoce con el nombre de método del "vecino más cercano" (nearest neighbor). Si en lugar de usar el caso más cercano se utilizan los k casos más similares, entonces se habla de los k-vecinos más cercanos (k Nearest Neighbors, kNN) y la clase asignada a la nueva instancia será la más común entre las k instancias más cercanas encontradas en la base de casos (Mitchell, 1997).

1.2 Árboles de decisión

El aprendizaje usando árboles de decisión es un método para aproximar funciones. Los árboles de decisión pueden también representarse como conjuntos de reglas IF-THEN. Un árbol de decisión clasifica las instancias ordenándolas de la raíz a las hojas. Cada nodo interior del árbol especifica una prueba de algún atributo y las hojas son las clases en las cuales se clasifican las instancias, cada rama descendiente de un nodo interior

corresponde a un valor posible del atributo probado en ese nodo. Un árbol de decisión representa una disyunción de conjunciones sobre los valores de los atributos. Así, cada rama, de la raíz a un nodo hoja, corresponde a una conjunción de atributos y el árbol en sí, a una disyunción de estas conjunciones.

La familia de algoritmos ID3 (Quinlan, 1986) es el paradigma de los métodos para descubrir reglas usando árboles de decisión, a pesar de esto, él tiene algunas limitaciones. Las más típicas son cuando el conjunto de entrenamiento tiene ruidos en los datos o cuando la cantidad de ejemplos es demasiado pequeña, lo cual puede llevar a la aparición del fenómeno conocido como sobreestimación (overfit). Entre las vías para enfrentar la sobreestimación están: detener el crecimiento del árbol antes de alcanzar un nivel de clasificación perfecta, o podar el árbol sobreestimado. Otro aspecto problemático en el uso de ID3 es que este algoritmo trabaja con atributos de dominio discreto, sin embargo, en muchas ocasiones es necesario usar atributos continuos para describir los objetos del dominio. Para resolver este problema basta con discretizar los atributos continuos.

Una variante para la solución de estas limitaciones es el algoritmo C4.5 (Quinlan, 1993), que usa puntos de corte e introduce varias medidas para evitar el sobre entrenamiento, en particular los criterios de parada de la división y de poda del árbol. El algoritmo C4.5 se basa en la utilización del criterio razón de ganancia (gain ratio). De esta manera se consigue evitar que las variables con mayor número de posibles valores salgan beneficiadas en la selección. Además el algoritmo C4.5 incorpora una poda del árbol de clasificación una vez que este ha sido inducido. La poda está basada en la aplicación de una prueba de hipótesis que trata de responder a la pregunta de si merece la pena expandir o no una determinada rama.

Otros árboles de decisión son el CHAID (Chi Square Automatic Interaction Detector) en el que la segmentación ocurre siguiendo criterios chi-cuadrados. En cada paso, CHAID selecciona la variable independiente (predictora) que tenga una interacción más fuerte (acorde al test chi-cuadrado) con la variable dependiente. Las categorías de los predictores se unen si no existen entre ellas diferencias significativas con respecto a la variable dependiente.

Otro de los métodos de creación de árboles es el CRT (Classification and Regression Tree). CRT divide los casos en segmentos que son lo más homogéneos posibles con respecto a la variable dependiente. Un nodo terminal es puro si en él todos los casos tienen el mismo valor de la variable independiente. Este método tiene la peculiaridad de que produce sólo árboles binarios. *

1.3 Redes bayesianas

Una red bayesiana es un modelo gráfico probabilístico que representa un conjunto de variables y sus dependencias probabilísticas. Las redes bayesianas permiten declarar supuestos de independencia condicionales que son aplicados a subconjuntos de variables. Son representadas por un gráfico acíclico dirigido, donde cada variable se representa por un nodo de la red, y de ella se especifican dos tipos de información:

(1) la estructura de dependencias condicionales que son los arcos de la red

(2) las distribuciones de probabilidad correspondientes.

Una red bayesiana puede calcular la distribución de probabilidad para cualquier subconjunto de variables de la red dado los valores o distribuciones de las variables restantes (Mitchell, 1997). Cuando no se conocen todos los valores de las variables en el conjunto de entrenamiento, el aprendizaje con una red bayesiana puede ser más difícil.

Este tipo de clasificador no es muy sensible a los cambios de sus parámetros, ya que se basa en información de toda la base, lo cual hace que pequeños cambios en la base no sean necesariamente significativos. (Chavez, 2008).

1.4 Máquinas de Soporte Vectorial

Las máquinas de soporte vectorial, también conocidas como máquinas de vectores de soporte (Support Vector Machine, SVM), son una técnica de aprendizaje supervisado muy interesante que se desarrolló en los últimos años (Vapnik, 1995), partiendo de la teoría de aprendizaje estadístico y basada en el principio de minimización de riesgo estructural. Se usa mucho tanto para resolver problemas de clasificación, como para regresión.

Concretamente, fundamenta las decisiones de clasificación, no basadas en todo el conjunto de datos, sino en un número finito y reducido de casos, que constituyen los "vectores soporte". Puede dividirse en SVM lineal y no lineal, este último en dependencia de diferentes funciones núcleo (kernel).

Algunas de las funciones núcleo más comúnmente usadas son la polinomial y la gaussiana de base radial o también conocida como función de base radial (Radial Basic Function, RBF), que se muestran en las ecuaciones E.1 y E.2 respectivamente.

Polinomial: $k(x, x') = (x \cdot x')^d$ E.1

Gaussiana de base radial: $k(x, x') = exp\left(\frac{\|x - x'\|}{2\sigma^2}\right)$ E..2

1.5 Redes neuronales artificiales

Una red neuronal es un modelo computacional que pretende simular el funcionamiento del cerebro a partir del desarrollo de una arquitectura que toma rasgos del funcionamiento de este órgano sin llegar a desarrollar una réplica del mismo (Bello et al., 2001). Es una herramienta matemática para la modelación de problemas, que permite obtener las relaciones funcionales subyacentes entre los datos involucrados en problemas de clasificación, reconocimiento de patrones, regresiones, etc. Este tipo de método se considera como un excelente aproximador de funciones, esencialmente no lineales, siendo capaces de aprender las características relevantes de un conjunto de datos, para luego reproducirlas en entornos ruidosos o incompletos (Wolpert, 1992).

Los modelos de redes neuronales son especificados por la función que caracteriza los nodos (modelo de la neurona), la topología de la red (estructura y tipo de enlaces) y las reglas o algoritmos de aprendizaje (método de ajustar los pesos).

El modelo de la neurona define el comportamiento de la misma al recibir una entrada para producir una respuesta. La topología no es más que la organización o arquitectura del conjunto de neuronas que la forman; esta organización comprende la distribución espacial de las mismas y los enlaces entre ellas. Los algoritmos de entrenamiento constituyen métodos que se aplican sobre los modelos de red para ajustar sus pesos y

obtener un comportamiento determinado. Con frecuencia los algoritmos de entrenamiento son caracterizados por la clase de topologías sobre las que se aplica, los tipos de parámetros libres que afecta (pesos de las conexiones entre neuronas, parámetros del algoritmo de entrenamiento, la topología misma de la red, etc.) y la regla de modificación de los mismos (Bonet, 2008).

En los últimos años se han producido una amplia variedad de arquitecturas de redes neuronales, encontrándose entre las más utilizadas, las redes multicapa de alimentación hacia adelante (Feed-Forward Neuronal Networks, FFN), las cuales se distinguen porque sus neuronas están conectadas a manera de grafo acíclico dirigido (todos los arcos hacia adelante). Las redes Multi Layer Perceptron (MLP) constituyen un ejemplo genérico de las redes FFN, y se encuentran formadas por un conjunto de capas de neuronas ordenadas secuencialmente. Primero una capa de entrada, luego un conjunto de capas intermedias denominadas capas ocultas y por último una capa de salida. Las MLP usando neuronas ocultas con funciones no lineales, son capaces de aproximar cualquier tipo de función continua y brindar excelentes resultados en las tareas de clasificación (Salazar, 2005).

1.6 Análisis discriminante

El análisis discriminante es quizás el primero que surgió entre todos los métodos de clasificación. Fue enunciado pos Sir Ronald Fisher en 1936. Es una técnica matemática que ayuda a identificar las características que diferencian (discriminan) a dos o más grupos y a crear una función (generalmente lineal) capaz de distinguir con la mayor precisión posible a los miembros de uno u otro grupo.

Obviamente, para llegar a conocer en qué se diferencian los grupos se necesita disponer de la información (cuantificada en una serie de variables) en las que se supone que se diferencian.

El análisis discriminante es una técnica estadística capaz de determinar cuáles variables permiten diferenciar a los grupos y cuántas de estas variables son necesarias para alcanzar la mejor clasificación posible. La pertenencia a los grupos, conocida de antemano, se utiliza como variable dependiente (una variable categórica con tantos valores discretos como grupos). Las variables en las que suponemos que se diferencian los grupos se

utilizan como variables independientes o variables de clasificación (también llamadas variables discriminantes). Ellas deben ser variables cuantitativas continuas o, al menos, admitir un tratamiento numérico ordinal.

El objetivo último del análisis discriminante es encontrar la combinación lineal de las variables independientes que mejor permite diferenciar (discriminar) a los grupos. Una vez encontrada esa combinación (la función discriminante) podrá ser utilizada para clasificar nuevos casos.

La función discriminante puede ser también cuadrática o polinomial de manera general. La limitante del método es que la forma de la función la decide el investigador a priori.

El análisis discriminante es una técnica multivariada que es capaz de aprovechar las relaciones existentes entre una gran cantidad de rasgos o variables independientes para maximizar la capacidad de discriminación.

1.7 Regresión logística

La regresión logística es un instrumento estadístico de análisis multivariado, de uso tanto explicativo como predictivo. Resulta útil su empleo cuando se tiene una variable dependiente dicotómica (un atributo cuya ausencia o presencia se ha puntuado con los valores cero y uno, respectivamente) y un conjunto de variables predictoras o independientes, que pueden ser cuantitativas (que se denominan covariables o covariadas) o categóricas. En este último caso, se requiere que sean transformadas en variables "dummy", es decir variables simuladas.

El propósito del análisis consiste en predecir la probabilidad de que a alguien le ocurra cierto "evento" (estar desempleado =1 o no estarlo = 0, enfermo = 1 o sano = 0). Puede además determinar cuáles variables pesan más para aumentar o disminuir la probabilidad de que a alguien le suceda el evento en cuestión. Esta asignación de probabilidad de ocurrencia del evento a un cierto sujeto, así como la determinación del peso que cada una de las variables dependientes en esta probabilidad, se basan en las características que presentan los sujetos a los que, efectivamente, les ocurren o no estos sucesos.

Por ejemplo, la regresión logística tomará en cuenta los valores que asumen en una serie de variables (edad, sexo, nivel educativo, posición en el hogar, origen migratorio, etc.) los sujetos que están efectivamente desocupados (=1) y los que no lo están (=0). En base a ello, predecirá a cada uno de los sujetos, independientemente de su estado real y actual, una determinada probabilidad de ser desocupado (es decir, de tener valor 1 en la variable dependiente). Si alguien es un joven no jefe de hogar, con baja educación y de sexo masculino y origen emigrante (aunque esté ocupado) el modelo le predecirá una alta probabilidad de estar desocupado (puesto que la tasa de desempleo del grupo así definido es alta), generando una variable con esas probabilidades estimadas, y procederá a clasificarlo como desocupado en una nueva variable, que será el resultado de la predicción. Y además, analizará cuál es el peso de cada uno de estas variables independientes en el aumento o la disminución de esa probabilidad. Por ejemplo, cuando aumenta la educación disminuirá en algo la probabilidad de ser desocupados. En cambio, cuando el sexo pase de 0 = mujer a 1 = varón, aumentará en algo la probabilidad de desempleo porque la tasa de desempleo de los jóvenes de sexo masculino es mayor que la de las jóvenes mujeres. El modelo, obviamente, estima los coeficientes de tales cambios (Chitarroni, 2002).

Por sus características, los modelos de regresión logística permiten dos finalidades:

- Cuantificar la importancia existente entre cada una de las covariables y la variable independiente, lo que lleva implícito también clarificar la existencia de interacción y confusión entre covariables respecto a la variable dependiente (es decir, conocer los odds ratio[1] para cada covariable)
- Clasificar individuos dentro de las categorías (presente/ausente) de la variable dependiente, según la probabilidad que tenga de pertenecer a una de ellas dada la presencia de determinadas covariables.

[1] Traducida al castellano con múltiples nombres como: razón de productos cruzados, razón de disparidad, razón de predominio, proporción de desigualdades, razón de oposiciones, oposición de probabilidades contrarias, cociente de probabilidades relativas, oportunidad relativa.

La regresión logística sólo resuelve problemas de clasificación binarios. Si el problema fuese más general, entonces se puede aplicar un modelo más general basado en los mismos principios, denominado regresión multinomial.

A pesar de los muchos estudios realizados hasta la actualidad en relación a los clasificadores no existe uno por excelencia, por lo que se hace difícil seleccionar el clasificador que logre encontrar una mejor frontera de decisión para separar las clases. En la búsqueda de mejores métodos de clasificación aparece una tendencia a combinar varios clasificadores en el mismo problema. En esto último se basan los algoritmos llamados multiclasificadores, utilizar varios clasificadores y combinar sus diferentes salidas (Polikar, 2006) con el objetivo de alcanzar un mejor resultado.

Dietterich (Dietterich, 2000) sugiere tres tipos de razones por las cuales un sistema multiclasificador puede ser mejor que un clasificador simple. La primera es estadística, pues si efectivamente por cada clasificador tenemos una hipótesis, la idea de combinar estas hipótesis, da como resultado una hipótesis que puede no ser la mejor, pero al menos evita seleccionar la peor de ellas. La segunda justificación es computacional, ya que algunos algoritmos ejecutan búsquedas que pueden llevar a diferentes óptimos locales: cada clasificador comienza la búsqueda desde un punto diferente y termina cercano al óptimo. Existe la expectativa de que alguna vía de combinación puede llevar a un clasificador con una mejor aproximación. La última justificación es figurativa ya que es posible que el espacio de hipótesis considerado no contenga la hipótesis óptima; pero la aproximación de varias fronteras de decisión puede dar como consecuencia una nueva hipótesis fuera del espacio inicial y que se aproxime más a la óptima.

3. Modelos de combinación de clasificadores

La combinación de clasificadores es en la actualidad un área activa de investigación en el aprendizaje automatizado y el reconocimiento de patrones. Se han publicado numerosos estudios, teórico y empíricos que demuestran las ventajas del paradigma de combinación de clasificadores por encima de los modelos individuales (Kuncheva, 2004).

Existen varias formas en las cuales se pueden construir multiclasificadores. Hay una serie de algoritmos desarrollados, algunos para problemas generales como bagging y boosting y otros para problemas específicos, pero todos tienen como partes fundamentales: la selección de los clasificadores de base y la elección de la forma de combinar las salidas (Bonet, 2008).

Entre los modelos más populares que combinan clasificadores están vote, bagging, boosting y stacking.

2.1 Bagging y Boosting

Bagging (Breiman, 1996) es uno de los primeros algoritmos de multiclasificadores, se basa en crear diferentes conjuntos de entrenamiento, extraídos del conjunto inicial de manera aleatoria y con remplazo, con lo cual asegura la diversidad. Este modelo necesita la selección de un modelo de clasificador inestable, o sea, un modelo que con pequeños cambios obtenga valores diferentes (Witten and Frank, 2005). Además usa un único modelo de clasificador y la combinación de los clasificadores resultantes se realiza con la técnica de voto mayoritario, ver figura 1. (Francia and García, 2006).

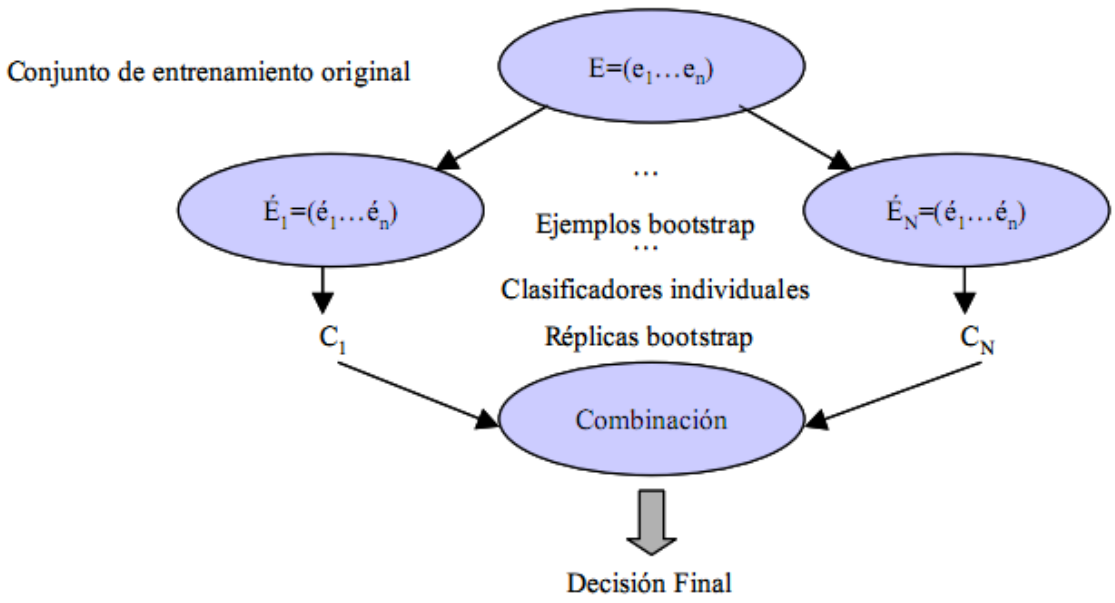

Figura 1. Esquema del método Bagging

Bagging puede ser aplicado en métodos de aprendizaje con predicciones numéricas, en los cuales las salidas individuales, que son números reales, son promediadas. Bagging usa para combinar las salidas voto mayoritario pero, en aras de lograr un mayor rendimiento, comúnmente se usa un estimado de probabilidad en lugar de una salida concreta, estas probabilidades estimadas por los modelos se promedian (Witten and Frank, 2005) y se le asigna la clase más probable. Una de las variaciones de este algoritmo se conoce con el nombre de "ramdom forests" (Breiman, 2001) que es construido con árboles de decisión como modelo de clasificador, y para crear los conjuntos de entrenamiento, usa el método de muestreo con remplazo como en bagging, o puede usar subconjuntos de rasgos.

Boosting (Schapire, 1990) es parecido a bagging porque usa el método de crear bases de entrenamiento aleatorias con reemplazo, a partir de la base original y un único modelo de clasificación para los clasificadores de base. Sin embargo, este algoritmo se realiza de manera secuencial, donde los clasificadores se van entrenando uno detrás del otro porque usan información del anterior. Otra diferencia es que boosting le da un peso al modelo por su rendimiento, en lugar de dar peso igual a todos los modelos. El remplazo se realizada estratégicamente de forma que los casos mal clasificados tienen mayor probabilidad, que los bien clasificados, de pertenecer al conjunto de entrenamiento del siguiente clasificador del sistema. Existen muchas variantes que utilizan la idea de boosting, siendo AdaBoost hoy en día la más utilizada (Freund and Schapire, 1997). Es una versión más general que se ha dividido en AdaBoost.M1 y AdaBoost.R, para problemas de clasificación y de regresión respectivamente. AdaBoost usa como método de combinación el voto mayoritario pesado.

2.2 Stacking

Stacking (Wolpert, 1992) es un método diferente a los anteriores pues la diversidad la busca con el empleo de diversos modelos de clasificación. Es menos utilizado que bagging y boosting, ya que es difícil de analizar teóricamente. Stacking combina múltiples clasificadores generados por diferentes algoritmos para un mismo conjunto de datos en una primera fase.

Para combinar las salidas no utiliza voto mayoritario, sino que introduce un metaclasificador, que aprende la relación entre las salidas de los clasificadores de base y la clase original. Esta es la segunda fase. Este metaclasificador tiene como base de entrenamiento un nuevo conjunto de instancias formadas a partir del conjunto de entrenamiento inicial para los clasificadores de base, donde por cada instancia del conjunto de entrenamiento se tiene ahora un vector de rasgos compuesto por las clases de salida de cada clasificador de base y como clase, la original de la instancia. Puede ser aplicado a predicciones numéricas (Witten and Frank, 2005).

La figura 2 muestra un ejemplo de un método Stacking.

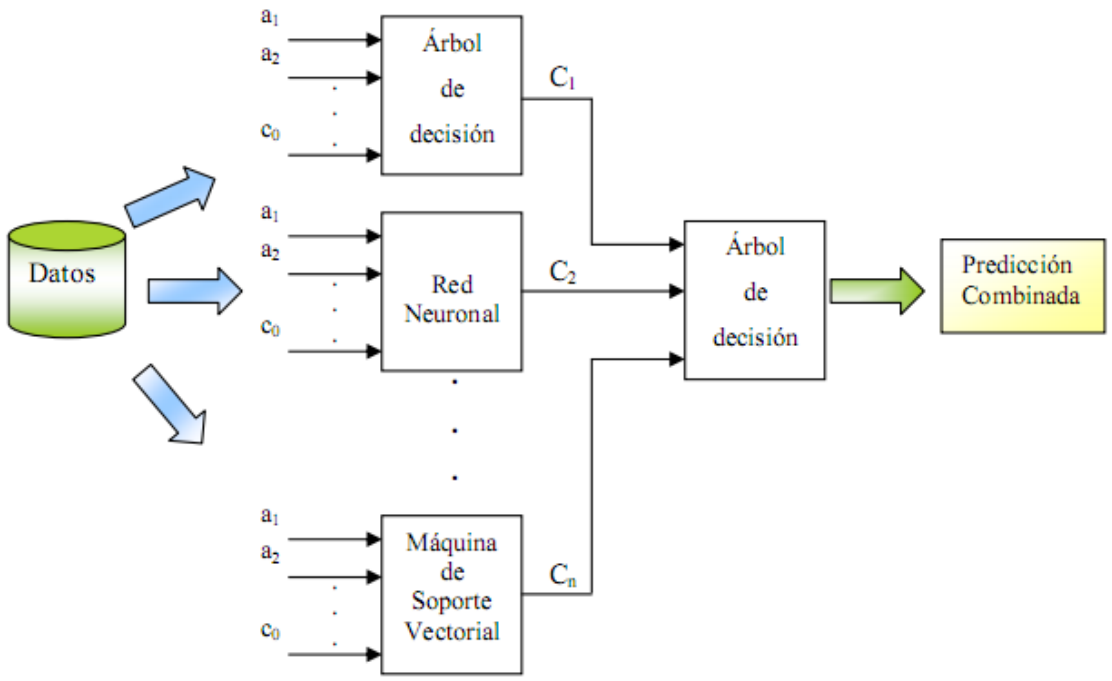

Figura 2. Esquema del método Stacking

2.3 Método de voto de clasificadores

El método de voto de clasificadores (conocido en WEKA como Vote) al igual que Stacking busca la diversidad con la utilización de diferentes modelos de clasificación como clasificadores de base. Las salidas estos clasificadores están dadas por vectores con una distribución de probabilidad para cada una de las clases. Vote combina estas

probabilidades, promediándolas, por voto mayoritario, tomando el mínimo, el máximo o la media.

Para comprender mejor algunas de estas funciones de combinación de salidas consideremos L clasificadores $(C_1, C_2, C_3, \ldots, C_L)$ para un problema de clasificación de m clases $(w_1, w_2, w_3, \ldots, w_m)$, cada clasificador tiene como salida un vector con una distribución de probabilidad para cada una de las clases. Un clasificador C_j tiene para la clase w_i un estimado de probabilidad $P(C_j,w_i)$, $i=1,2..,m$. De esto último se tiene que:

$$\sum_{i=1}^{m} P(C_j, w_i) = 1 \qquad \text{E..3}$$

El estimado de probabilidad de un conjunto de clasificadores para cada una de las clases de un caso X se puede representar por la matriz MP.

$$MP(X) = \begin{bmatrix} P(C_1,w_1) & \ldots & P(C_1,w_i) & \ldots & P(C_1,w_m) \\ \ldots & & \ldots & & \ldots \\ P(C_j, w_1) & \ldots & P(C_j, w_i) & \ldots & P(C_j, w_m) \\ \ldots & & \ldots & & \ldots \\ P(C_L,w_1) & \ldots & P(C_L,w_i) & \ldots & P(C_L,w_m) \end{bmatrix}$$

Salida del clasificador C_L

Estimado de los clasificadores $C_1, C_2, \ldots, C_L$ para la clase w_i

A partir de la matriz MP se pueden definir las siguientes funciones de combinación de salida:

$$Promedio_{w_i} = \frac{\sum_{j=1}^{L} MP(j,i)}{L}, \qquad donde\ L\ es\ el\ total\ de\ clasificadores \qquad \text{E. 4}$$

$$Máximo_{w_i} = \frac{V(w_i)}{\sum_{i=1}^{m} V(w_i)}, \qquad donde \qquad V(w_i) = max_{j=1}^{L} MP(j,i) \qquad \text{E. 1}$$

La última función de combinación de salida que será analizada se conoce como voto mayoritario. Para ello primeramente se construirá la matriz *MK* donde cada posición se calcula como:

$$MK_{j,i} = \begin{cases} 1 & si \quad MP(j,i) = max_{i=1}^{m} MP(j,i) \\ 0 & en\ otro\ caso \end{cases} \qquad \text{E. 6}$$

El voto para una clase sería:

$$Voto_{w_i} = \sum_{j=1}^{L} MK(j,i) \qquad \text{E. 7}$$

Luego la clase de voto mayoritario sería:

$$Clase\ de\ Voto\ Mayoritario = w_i, \quad donde \quad i = arg_i \left(max_{i=1}^{m} \sum_{j=1}^{L} MK(j,i) \right) \qquad \text{E. 8}$$

2.4 Métodos basados en rasgos

En la construcción de un multiclasificador, los clasificadores de base pueden ser construidos con subconjuntos de rasgos diferentes, lo cual es otra forma de buscar diversidad. La selección de rasgos tiene como objetivo una mayor eficiencia en los cálculos así como una mayor exactitud del multiclasificador. Existen muchos modelos de multiclasificadores que utilizan subconjuntos de rasgos diferentes como los descritos por Kuncheva (Kuncheva, 2004).

La elección de estos subconjuntos de rasgos puede ser realizada de diferentes vías. Entre ellas podemos mencionar la selección aleatoria (Random Selection), conocida también como método del subespacio aleatorio (ramdom subspace method), donde cada clasificador se construye sobre un subconjunto aleatorio de rasgos de tamaño *d*. Se han obtenido buenos resultados utilizando clasificadores con *d=n/2* rasgos, donde n es el

número total de rasgos. Con el método de selección aleatoria se han obtenido buenos resultados cuando hay información redundante la cual está dispersa por todos los rasgos en lugar de estar concentrada en un subconjunto (Skurichina, 2001, Ho, 1998).

Otro método propone combinar bagging con selección aleatoria (Random Selection). Se seleccionan B subconjuntos aleatorios de la base de entrenamiento y para cada uno de ellos se seleccionan R subconjuntos de rasgos. El multiclasificador consta de L=B+R clasificadores. Esta combinación de bagging con selección aleatoria busca alcanzar una mayor diversidad entre los clasificadores que con la utilización de cualquiera de los dos de manera independiente. Se ha demostrado que esta combinación es más ventajosa que si se usan por separado.

La elección de los subconjuntos también se puede realizar de manera no aleatoria. Un algoritmo utilizado es el conocido como "input decimation", donde al multiclasificador está compuesto por C clasificadores, donde C es igual al número de clases del problema. Cada clasificador tiene una clase favorita. Para encontrar el subconjunto de rasgos del clasificador C_i con clase favorita i, se calcula la correlación entre cada rasgo y la clase. Los rasgos correspondientes a la mayor correlación son elegidos como subconjunto para el clasificador C_i. Este método ha demostrado tener un mayor rendimiento que el método del subespacio aleatorio.

Otro método propone combinar bagging con selección aleatoria (Random Selection). Se seleccionan B subconjuntos aleatorios de la base de entrenamiento y para cada uno de ellos se seleccionan R subconjuntos de rasgos. El multiclasificador consta de L=B+R clasificadores. Esta combinación de bagging con selección aleatoria busca alcanzar una mayor diversidad entre los clasificadores que con la utilización de cualquiera de los dos de manera independiente. Se ha demostrado que esta combinación es más ventajosa que si se usan por separado.

Como se había mencionado anteriormente, la combinación de clasificadores es en la actualidad un área de investigación activa y de gran utilidad en el aprendizaje automatizado y el reconocimiento de patrones. Cada vez que se combinan clasificadores se necesita garantizar la diversidad entre ellos.

4. Selección de clasificadores de base para un modelo multiclasificador. Medidas de diversidad

La selección de los clasificadores de base es el primer paso a la hora de construir un multiclasificador.

Algunos paradigmas de combinar clasificadores usan el mismo modelo de clasificación, pero no existe evidencia de si esa estrategia es mejor que el uso de diferentes modelos (Kuncheva, 2004).

La diversidad entre los clasificadores de base es muy importante, ya que de esto dependerá en gran medida el resultado final del multiclasificador. La diversidad de los errores de los clasificadores puede dar una medida del mayor valor posible que se puede aspirar con la combinación de esos modelos. Como ya se discutió anteriormente, algunos multiclasificadores, como bagging y boosting, garantizan la diversidad utilizando diferentes conjuntos de bases de entrenamiento. Otros utilizan diferentes conjuntos de rasgos y otros diferentes clasificadores de base. En estos dos últimos casos no se garantiza una gran diversidad, por lo que es preciso el uso de algunas medidas estadísticas que permitan hacer estimación de cuán diversos son los clasificadores.

En los epígrafes que siguen se muestran medidas de diversidad descritas por Kuncheva en (Kuncheva and Whitaker, 2003) y algunas que han sido enunciadas por otros autores.

(Kuncheva and Whitaker, 2003) plantean que no hay una medida de diversidad involucrada en forma explícita en los métodos de generación de clasificadores, aunque asumen que la diversidad es el punto clave en cualquiera de los métodos. La elección de la medida a utilizar va a depender directamente de la cantidad de clasificadores a utilizar y de muchos otros aspectos que deben ser estudiados con detalle en investigaciones futuras. La mayoría de estas medidas son estadísticas.

Las medidas pueden ser clasificadas como: medidas en forma de pares (pairwise) y medidas para todo el conjunto (nonpairwise) (Kuncheva and Whitaker, 2003).

Las medidas en forma de pares se calculan por pares de clasificadores usando sus salidas, las cuales son binarias (0,1) que indica si la instancia fue correctamente clasificada o no por el clasificador.

A continuación se indica el resultado de dos clasificadores (C_i, C_j) para una instancia en cuanto si la clasificaron correctamente o no.

	C_j correcto (1)	C_j incorrecto (0)
C_i correcto (1)	*a*	*b*
C_i incorrecto (0)	*c*	*d*
a + b + c + d = 1		

Tabla 1. Tabla de clasificación entre los resultados de los clasificadores C_i y C_j

Si se suman para todas las instancias los valores de a, b, c, d entre el par de clasificadores (C_i, C_j) se obtendrá el siguiente resultado, a partir del cual se calculan las medidas en forma de pares:

	C_j correcto (1)	C_j incorrecto (0)
C_i correcto (1)	*A*	*B*
C_i incorrecto (0)	*C*	*D*
A +B + C + D = N		

Tabla 2. Tabla de clasificación entre los resultados de los clasificadores C_i y C_j

Donde *A* sería igual a la suma total de los valores de *a* para todas las instancias y así respectivamente con los valores de *B, C y D.* N es el número total de casos.

Un conjunto de L clasificadores produce L (L – 1)/2 pares de valores. Para obtener un único resultado habría que promediar estos valores.

Mientras que las medidas de diversidad que se basan en todo el conjunto consideran a todos los clasificadores a la vez y calculan un único valor de diversidad para todo el conjunto.

4.1 Medidas de diversidad en forma de pares (pairwise)

La figura 3 enuncia varias medidas de diversidad en forma de pares.

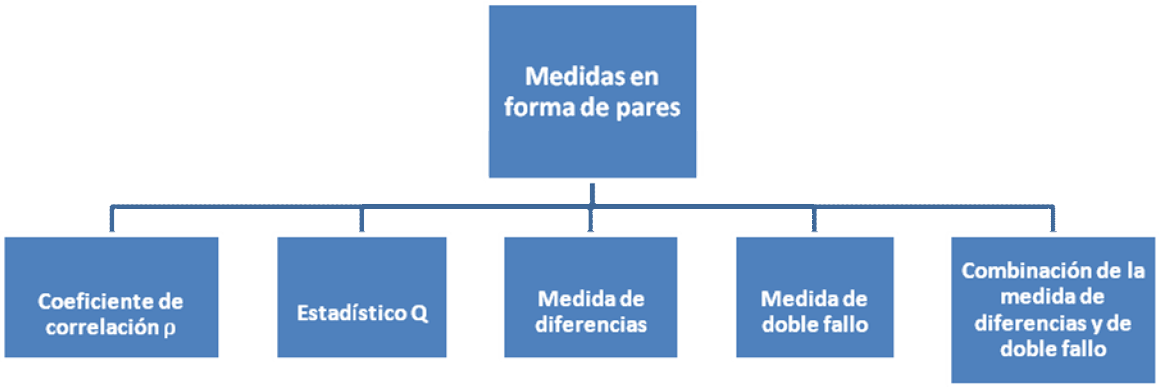

Figura 3. Medidas de diversidad en forma de pares

4.1.1 Coeficiente de correlación ρ

Entre las medidas de diversidad está el coeficiente de correlación el cual se calcula como,

$$\rho_{C_i,C_j} = \frac{A \cdot D - B \cdot C}{\sqrt{(A+B)\cdot(C+D)\cdot(A+C)\cdot(B+D)}} \quad , -1 \leq \rho \leq 1 \qquad \text{E. 9}$$

El coeficiente de correlación también puede ser calculado para pares de clasificadores que devuelven grados de pertenencia a cada clase. Para cada par de clasificadores van a existir c coeficientes de correlación, uno por cada clase. La medida final sería el promedio de los valores asociados a cada clase (Kuncheva, 2004). Mientras menor sea el valor de ρ en módulo, mayor será la diversidad.

4.1.2 El estadístico Q

El estadístico Q (Q Statistics) es otra de las medidas para pares de clasificadores. Se calcula de la siguiente forma:

$$Q_{ci,cj} = \frac{A \cdot D - B \cdot C}{A \cdot D + B \cdot C}, -1 \leq Q \leq 1$$ E. 10

Para un par de clasificadores estadísticamente independientes, su valor de $Q_{ci,cj}$ va a ser 0. En general, el valor que Q va a oscilar entre −1 y 1. Aquellos clasificadores que tienden a reconocer los mismos objetos correctamente tendrán un valor positivo de Q, y aquellos que comentan errores en diferentes objetos poseerán un valor negativo de Q.

Para cualquier par de clasificadores, los valores de ρ y Q tendrán el mismo signo y se puede probar que $|\rho| \leq |Q|$. (Kuncheva and Whitaker, 2003).

4.1.3 Medida de diferencias

La medida de diferencias (The Disagreement Measure) introducida por Skalak (Skalak, 1996.), es la más intuitiva de las medidas entre un par de clasificadores, y es igual a la probabilidad de que los dos clasificadores discrepen en sus predicciones. Mientras mayor sea su valor mayor será la diversidad.

$$D_{ci,cj} = \frac{B + C}{N}$$ E. 11

4.1.4 Medida de doble fallo

Otra de las medidas que se analizará se conoce como medida de doble fallo (The Double-Fault Measure) introducida por Giacinto y Roli (Giacinto and Roli, 2001.) considera el fallo de los dos clasificadores al mismo tiempo. Ruta y Gabrys (Ruta and Gabrys, 2001) definen a esta medida como una medida no-simétrica. Esto quiere decir que si se intercambian los unos con los ceros en los resultados de los clasificadores, el valor de la medida no va a ser el mismo. Esta medida está basada en el concepto de que es más importante conocer cuando errores simultáneos son cometidos que cuando ambos tienen clasificación correcta. Mientras menor sea el valor mayor será la diversidad.

$$DF_{Ci,Cj} = \frac{D}{N}$$ E. 12

4.1.5 Combinación de la medida de diferencias y medida de doble fallo

La última de las medidas para pares de clasificadores que es una combinación entre la medida de diferencias y la medida de doble fallo. Mientras mayor sea el valor de esta medida mayor será la diversidad entre los clasificadores.

$$R_{Ci,Cj} = \frac{D_{Ci,Cj}}{DF_{Ci,Cj}}$$ E. 13

4.2 Medidas de diversidad para todo el conjunto (nonpairwise)

La figura 4 enuncia varias medidas de diversidad grupales.

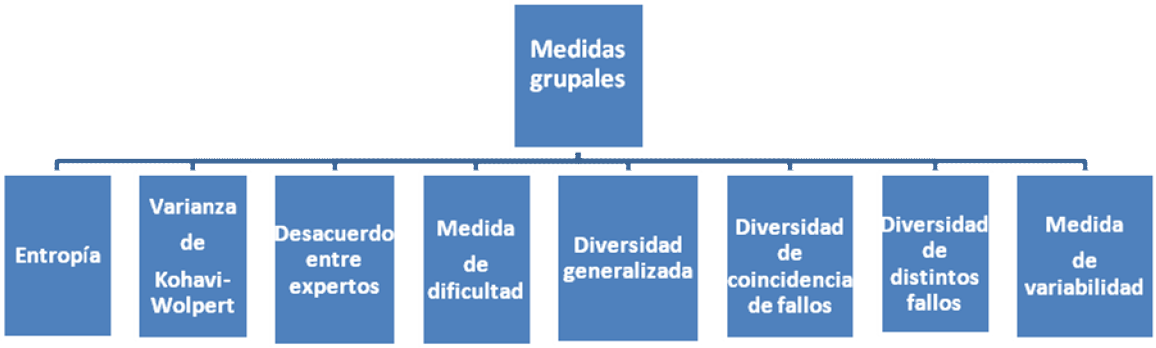

Figura 4. Medidas de diversidad grupales

4.2 1 Entropía

Entre estas medidas se encuentran la Entropía (The Entropy Measure) (Kuncheva and Whitaker, 2003), ésta se basa en la idea intuitiva de que en un conjunto de N casos y L clasificadores la mayor diversidad se obtendrá si L/2 de los clasificadores clasifican una

instancia correctamente y los otros L- L/2 la clasifican incorrectamente. Fue introducida por Cunningham y Carney (Cunningham and J. Carney, 2000).

$$E = \frac{1}{N} \cdot \frac{2}{L-1} \sum_{j=1}^{N} min\left\{\left(\sum_{i=1}^{L} y_{j,i}\right), \left(L - \sum_{i=1}^{L} y_{j,i}\right)\right\} \quad , y_{j,i} \in \{0,1\}, \quad 0 \leq E \leq 1 \qquad \text{E. 14}$$

Donde $y_{j,i}$ tendrá valor 1 si el clasificador i clasificó correctamente el caso j y 0 en caso contrario. Si E tiene valor 0 esto indica que no hay diferencia entre los clasificadores y un valor 1 indica la mayor diversidad.

4.2.2 Varianza de Kohavi-Wolpert

Otra de las medidas de diversidad es la conocida como Varianza de Kohavi-Wolpert (Kohavi-Wolpert Variance), fue inicialmente propuesta por Kohavi y Wolpert (Kohavi and Wolpert., 1996.). Esta medida es originada de la descomposición de la varianza del sesgo del error de un clasificador.

Kuncheva y Whitaker presentaron en (Kuncheva and Whitaker, 2003) una modificación para medir la diversidad de un ensamblado compuesto por clasificadores binarios, quedando la medida de diversidad como

$$KW = \frac{1}{NL^2} \sum_{j=1}^{N} Y(z_j)\left(L - Y(z_j)\right) \quad , \quad donde \;\; Y(z_j) = \sum_{i=1}^{L} y_{i,j} \qquad \text{E. 15}$$

Con esta medida la diversidad disminuye a medida que el valor de KW aumenta.

4.2.3 Medida de desacuerdo entre expertos

La medida de desacuerdo entre expertos (Measurement interrater agreement) (Fleiss, 1981) es otra de las medidas de diversidad que se basan en todo el conjunto. Se desarrolla como una medida de fiabilidad entre clasificadores. Puede usarse para medir el nivel de acuerdo dentro de un conjunto de clasificadores, por consiguiente esta también basada en el supuesto que un conjunto de clasificadores debe discrepar entre sí para ser diverso. La diversidad disminuye cuando el valor de k aumenta. El k se calcula por:

$$k = 1 - \frac{\frac{1}{L}\sum_{j=1}^{N} Y(Z_j)(L - Y(Z_j))}{N(L-1)p(1-p)}$$ E. 16

Donde el término de la derecha es la medida de concordancia de Kendall y p es la media de la exactitud de la clasificación individual, y se calcula como:

$$p = \frac{1}{N \cdot L} \cdot \sum_{j=1}^{N} \sum_{i=1}^{L} y_{j,i}$$ E. 17

4.2.4 Medida de dificultad

La medida de dificultad viene del estudio realizado por Hansen y Salamon (Hansen and Salamon., 1990.) (The Measure of "difficulty" ϴ), se calcula a través de la varianza de una variable aleatoria discreta que toma valores en el conjunto {0/L, 1/L, 2/L,..,1} y denota la probabilidad de que exactamente i clasificadores, hayan clasificado bien en todas las instancias.

Para conveniencia, θ suele ser escalada linealmente en el intervalo [0,1] tomando como p(1 − p) como el mayor valor posible, donde p es la precisión individual de cada clasificador. La diversidad del ensamblado aumenta con el decremento del valor de la medida de dificultad. Idealmente θ = 0, pero este es un escenario poco realista.

La intuición de esta medida puede ser explicada de la siguiente manera: Un ensamblado de clasificadores diverso tiene un valor pequeño de medida de dificultad, dado que cada muestra de entrenamiento puede al menos ser clasificada correctamente por una proporción de todos los clasificadores base, lo cual es más probable con una baja varianza de X. Mientras menor sea su valor mayor será la diversidad.

$$ϴ = Var(x)$$ E. 18

L es la cantidad de clasificadores.

4.2.5 Medida de diversidad generalizada

La medida de diversidad generalizada (Generalized Diversity) se enunció por Partridge y Krzanowski (Partridge and Krzanowski, 1997).

Sea *Y* una variable aleatoria que representa la proporción de clasificadores que clasificaron incorrectamente una muestra x □ R^n extraída aleatoriamente del conjunto de datos. Denotemos por p_i la probabilidad de que *Y*=i/L y p(i) la probabilidad de que i clasificadores extraídos de manera aleatoria fallen en clasificar correctamente un objeto x extraído aleatoriamente. Supongamos que dos clasificadores son tomados de forma aleatoria del ensamblado *D*, Partridge y Krzanowski exponen en su trabajo que la máxima diversidad es lograda cuando el uno de los dos clasificadores se equivoca en clasificar un objeto y el otro lo clasifica correctamente. En este caso la probabilidad de equivocarse los dos clasificadores es p(2)=0. Por otra parte argumentan que la mínima diversidad se lograría cuando el fallo de un clasificador es siempre acompañado con el fallo del otro, entonces la probabilidad de que los dos clasificadores fallen es la misma que la probabilidad de que un clasificador escogido de forma aleatoria falle.

$$GD = 1 - \frac{p(2)}{p(1)} \qquad p(1) = \sum_{i=1}^{L} \frac{i}{L} \times pi \qquad p(2) = \sum_{i=1}^{L} \frac{i(i-1)}{L(L-1)} \times pi \qquad \text{E. 19}$$

El valor de *GD* varía entre 0 y 1, siendo 0 la menor diversidad cuando p(2)=p(1) y 1 la mayor diversidad cuando p(2)=0 y L es la cantidad de clasificadores.

4.2.6 Medida de diversidad de coincidencia de fallos

Esta medida (Coincident Failure Diversity) se enuncia por Partridge y Krzanowski también (Partridge and Krzanowski, 1997), como una mejora a la medida anterior.

Esta medida está diseñada tal que tenga un valor mínimo 0 cuando todos los clasificadores siempre clasifiquen correctamente o cuando todos los clasificadores lo mismo clasifiquen correcta o incorrectamente al mismo tiempo. Su máximo valor 1 es alcanzado cuando todos los errores de clasificación son únicos, es decir cuando al menos un clasificador va a clasificar incorrectamente cualquier objeto aleatorio.

$$CFD = \begin{cases} 0 & si \;\; po = 1 \\ \frac{1}{1-po} \sum_{i=1}^{L} \frac{L-i}{L-1} \times pi & po < 1 \end{cases}$$ E. 20

p_o=1 cuando todos los clasificadores siempre son simultáneamente correctos o incorrectos.

p_i es el mismo término de la medida anterior y L es la cantidad de clasificadores.

4.2.7 Medida de diversidad de distintos fallos

Esta medida (Distintic Failure Diversity) fue igualmente enunciada por Partridge y Krzanowski (Partridge and Krzanowski, 1997.), como una mejora a la medida anterior, pues ahora se va a tener en cuenta todas las instancias donde los clasificadores no coinciden en las clases asignadas, es decir, se consideran las distintas posibilidades de fallo teniendo en cuenta las clases.

$$DFD = \begin{cases} 0 & si \;\; ti = 0 \\ \sum_{i=1}^{L} \frac{L-i}{L-1} \times ti & si \;\; ti < 0 \end{cases}$$ E. 21

Donde t_i es el número de i fallos ocurridos dividido por el total de fallos distintos ocurridos y L es la cantidad de clasificadores.

4.2.8 Medida de Variabilidad

Esta medida (The Measure of Variability*)* tiene en cuenta si las clases asignadas por los clasificadores en cada instancia son distintas o no. Mientras mayor sea su valor mayor será la diversidad.

$$Var = \frac{\sum_{y=1}^{p} a}{p} \text{ donde } a = \begin{cases} 0 & si \; E_1(y) = E_2(y) = \cdots = E_L(y) \\ 1 & e.o.c. \end{cases}$$ E. 22

Donde p es el total de instancias y Ei es la etiqueta (clase) asignada a la instancia "y", por el clasificador i.

5. Combinación de salidas

Otra parte importante en un multiclasificador es la descripción de cómo las salidas de cada clasificador se pueden combinar en una sola. Ésta puede dividirse en dos tipos: selección o fusión.

- La selección es la simple elección del "mejor" clasificador para una instancia determinada, o sea, dada una instancia, seleccionar del conjunto de clasificadores base cuál es el que dará la salida para ella.
- La fusión, por otro lado, se basa en combinar, mediante alguna función, las salidas de los diferentes clasificadores. Las salidas de los clasificadores pueden ser un valor concreto, o un vector de probabilidades con una probabilidad asociada a cada clase.

Una de las formas más conocidas de combinar las clases es el voto mayoritario, ya sea simple o pesado. Aquí cada clasificador le da un voto a las clases acorde con su resultado y finalmente se selecciona la clase con mayor cantidad de votos. Las figuras 5 y 6 muestran estas ideas gráficamente mediante un ejemplo con tres clasificadores.(Francia and García, 2006).

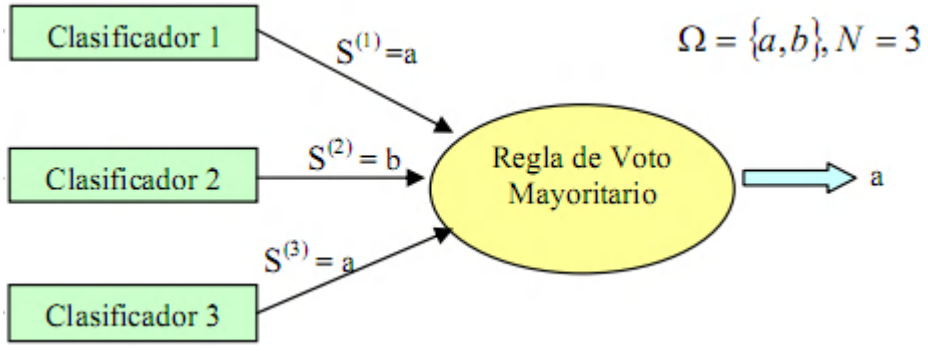

Figura 5. Ejemplo del método de voto mayoritario

En el voto mayoritario simple, todos los clasificadores tienen la misma importancia. No ocurre así en la versión pesada.

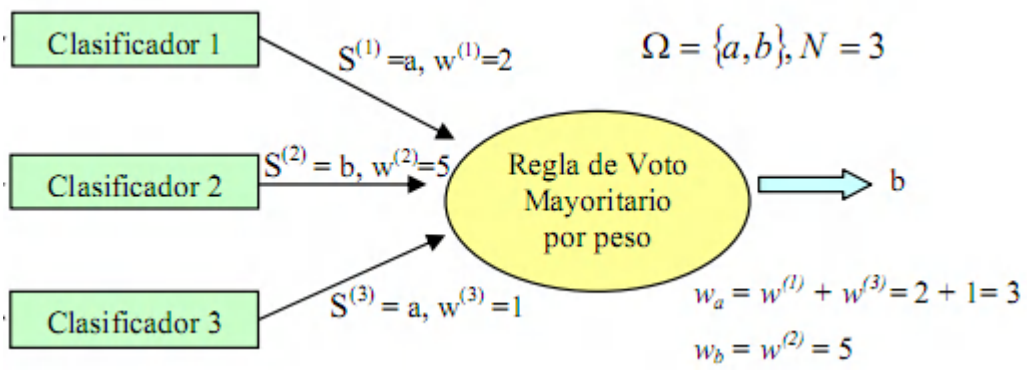

Figura 6. Ejemplo del método de voto mayoritario pesado

La mayoría de los clasificadores dan como salida un vector de probabilidades, que se conoce como distribución de probabilidad, donde se da una probabilidad para cada clase, de forma que la salida final del clasificador es aquella clase con mayor probabilidad asociada. Utilizando la distribución de probabilidad de los clasificadores de base pueden utilizarse otras variantes para la combinación de las salidas, como pueden ser la elección de la clase que tenga la menor probabilidad, la de mayor, la de mayor producto, la de probabilidad media, etc. (Kuncheva, 2004).

Otra forma de combinar las salidas es la introducción de un clasificador que aprende la relación entre las salidas de los clasificadores y la clase real. Este clasificador es conocido en la literatura como metaclasificador.

6. Evaluación en la clasificación

Para evaluar la clasificación, existen una serie de medidas que se calculan a partir de los resultados de las predicciones del modelo cuando se prueban en un conjunto de datos que no intervienen en el entrenamiento.

Entre estas medidas podemos destacar: el error, la exactitud, la razón de verdaderos positivos y la razón de falsos positivos.

Si bien con las medidas antes mencionadas se puede determinar qué tan bueno es un modelo de clasificación, la forma de dividir los datos en conjunto de entrenamiento y prueba es también muy importante. Existen diferentes técnicas para esto, como es el

método hold-out el cual reserva una cierta cantidad de casos para probar y usa el resto para entrenamiento, por lo general se entrena con 2/3 de los datos y utiliza 1/3 para prueba. Tiene como dificultad que si hay una cantidad de datos pequeña los ejemplos pueden no ser representativos ya que puede ocurrir que haya pocas o ninguna instancia de algunas clases, también puede dar una solución "buena" a partir de haber hecho una división de la base muy oportuna.

Con el objetivo de buscar una mayor precisión en el método hold-out se creó el hold-out repeated que es repetir hold-out varias veces creando de manera aleatoria los conjuntos de entrenamiento y prueba, los errores de cada iteración se promedian para dar el resultado final.

Otra alternativa es usar el conjunto de datos de entrenamiento para prueba, esto nos puede traer el problema conocido como overfitting o sobreentrenamiento, además de que puede llegar a conclusiones apresuradas pues no se está teniendo en cuenta cómo reacciona el modelo ante casos que no ha visto antes, que es lo que le daría la medida de generalización.

El método de validación cruzada con k subconjuntos (k-fold cross-validation) es uno de los más usados, este método se basa en dividir la base en k partes iguales. Se realizan k entrenamientos del modelo, tomando siempre una parte para prueba y las restantes para entrenamiento, se mide el error con los ejemplos de prueba. Tiene la ventaja que todos los ejemplos de la base de casos son, eventualmente usados para ambos procesos, entrenamiento y prueba. Como inconveniente podemos señalar que en grandes volúmenes de datos la validación es muy lenta (Efron and Tibshirani, 1993).

7. Estudio de un caso

Se quiere resolver un problema real que consiste en clasificar el posible padecimiento de hipertensión en niños de la ciudad de Santa Clara, Cuba, entre ocho y once años de edad. Los datos que se utilizaron en esta investigación, fueron suministrados por el proyecto PROCDEC de la Universidad Central "Marta Abreu" de Las Villas, Santa Clara, Cuba (González, 2010).

La muestra consta de un total de 795 pacientes, contiene datos de niños de cuatro escuelas primarias de la ciudad de Santa Clara. A continuación se muestran los resultados de aplicar varios modelos de clasificación al problema presentado. Los cálculos se hicieron con ayuda del software WEKA (*Waikato Environment for Knowledge Analysis*)[2] , con la opción "Percentage Split" y los parámetros por defecto del software, en esta opción una parte de los datos se utilizan para el entrenamiento de los clasificadores por eso el total de casos clasificados no coincide con el total de la base.

	Normotensos	***Hipertensos***
Sanos	122	24
Enfermos	37	88
Exactitud	**77.49%**	

Tabla 3. Resultados del clasificador adtrees

	Normotensos	***Hipertensos***
Sanos	104	42
Enfermos	58	67
Exactitud	**63.09%**	

Tabla 4. Resultados del clasificador IB1

	Normotensos	***Hipertensos***
Sanos	128	18
Enfermos	35	90
Exactitud	**80.44%**	

Tabla 5. Resultados del clasificador logistic

	Normotensos	***Hipertensos***
Sanos	125	21
Enfermos	35	90
Exactitud	**79.33%**	

Tabla 6. Resultados del clasificador naivesbayes

[2]http://www.cs.waikato.ac.nz/ml/weka/

	Normotensos	*Hipertensos*
Sanos	124	22
Enfermos	44	81
Exactitud	**75.64%**	

Tabla 7. Resultados del clasificador OneR

	Normotensos	*Hipertensos*
Sanos	119	27
Enfermos	50	75
Exactitud	**71.58%**	

Tabla 8. Resultados de una red Multilayer Perceptron

Una vez obtenidos estos resultados se calcularon algunas de las medidas de diversidad anteriormente explicadas, las cuales arrojaron el conjunto de clasificadores más diversos, teniendo en cuenta sólo las combinaciones de tres que se pueden formar con estos clasificadores. Dentro de este conjunto se encuentran los clasificadores: ADTrees, Logistic y NaivesBayes. El 9 que aparece en la columna "Cant" significa que este conjunto estuvo entre los 3 mejores en 9 medidas de las 13.

Resultado de las Medidas													
R	p	D	Q	DF	KW	K	E	DIF	GD	CFD	DFD	Var	Cant
2.70	0.25	0.30	0.51	0.11	0.10	0.22	0.45	0.84	0.93	0.85	0.56	0.45	9

Tabla 9. Resultados de las medidas de diversidad

Luego se utilizó el método de voto de clasificadores (Vote) el cual, al igual que Stacking, busca la diversidad con la utilización de diferentes modelos de clasificación como clasificadores de base. Las salidas de estos clasificadores están dadas por vectores con una distribución de probabilidad para cada una de las clases. Vote combina estas probabilidades, promediándolas, por voto mayoritario, tomando el mínimo, el máximo, la media o el producto de ellas, se utilizó este último criterio.

Los clasificadores bases usados fueron los más diversos (mencionados anteriormente) según las medidas de diversidad. A continuación se muestran los resultados.

	Normotensos	*Hipertensos*
Sanos	129	7
Enfermos	31	104
Exactitud	**85.97%**	

Tabla 10. Resultados del multiclasificador Vote con combinación más diversa

Se presenta la aplicación de varios clasificadores a la solución del diagnóstico de la hipertensión arterial infantil en la ciudad de Santa Clara. En todos ellos la exactitud fue inferior al 80%.

La combinación de los clasificadores más diversos, usando los resultados de las medidas de diversidad, dio resultados positivos. El porcentaje de casos bien clasificados se elevó a casi un 86%. Con esto se demuestra la ventaja de la combinación de clasificadores diversos.

Conclusiones

Los clasificadores tienen innumerables aplicaciones en muchas áreas. La combinación apropiada de dos o más de ellos puede proporcionar una predicción más precisa, exacta y eficiente que el uso de un único clasificador.

Este trabajo constituye un estudio organizado y ordenado de la combinación de clasificadores supervisados. Recoge una actualización de los diversos métodos y sistemas desarrollados en los últimos años. El propósito de realizar un estudio de las diferentes técnicas existentes en esta área para que sirva de punto de partida para estudiantes e investigadores que deseen conocer más sobre el tema queda completamente cumplido.

Se presentan las ideas generales de los clasificadores más frecuentemente utilizados en la solución de problemas reales: algoritmos basados en casos, árboles de decisión con diferentes criterios de segmentación, redes bayesianas, máquinas de soporte vectorial, redes neuronales artificiales, el clásico análisis discriminante y la regresión logística.

Se realiza un análisis crítico de los sistemas multiclasificadores reportados en la literatura, entre los que pueden mencionarse: bagging, boosting, stacking y voto mayoritario entre otros.

Uno de los aspectos medulares lo constituye la presentación y discusión crítica de las medidas de diversidad que aprueban o no la combinación de un conjunto de clasificadores, pues es frecuente combinar clasificadores sin usar medidas de diversidad que aseguren el éxito del multiclasificador.

Se comentan además las ideas esenciales del problema de la combinación de las salidas de los clasificadores.

La presentación de un caso de estudio: diagnóstico de la hipertensión arterial en niños de edad escolar en Santa Clara muestra, a manera de ejemplo, las bondades del uso de los multiclasificadores.

A pesar de los grandes avances en esta área, aún queda mucho por hacer. Esta monografía debe servir de punto de partida para la realización de investigaciones futuras que resuelvan o mejoren algunas de las dificultades aquí planteadas.

Referencias Bibliográficas

BELLO, R., GARCÍA, Z., GARCÍA, M. M. & LOBATO, A. R. 2001. *Aplicaciones de la IA,* Santa Clara.

BONET, I. 2008. *Modelo para la clasificación de secuencias, en problemas de la bioinformática, usando técnicas de inteligencia artificial.* Doctorado, Universidad Central "Martha Abreu" de las Villas.

BREIMAN, L. 1996. Bagging predictors. *Machine Learning,* 24**,** 123-140.

BREIMAN, L. 2001. Random Forests. *Machine Learning,* 45**,** 5–32.

CHAVEZ, M. D. C. 2008. *MODELOS DE REDES BAYESIANAS EN EL ESTUDIO DE SECUENCIAS GENÓMICAS Y OTROS PROBLEMAS BIOMÉDICOS.* Doctorado, Universidad Central "Martha Abreu" de las Villas.

CHITARRONI, H. 2002. *La regresión logística* [Online]. Buenos Aires, Argentina: Facultad de Ciencias Sociales. Universidad del Salvador. Available: www.salvador.edu.ar/csoc/idicso/docs/aephc1.pdf [Accessed 2008 9 /ene/].

CUNNINGHAM, P. & J. CARNEY 2000. Diversity versus Quality in Classification Ensembles Based on Feature

Selection, in Machine Learning:. *ECML , R. López de Mántaras and E. Plaza, Editors. 2000, Springer Berlin / Heidelberg.***,** p. 109-116.

DIETTERICH, T. G. 2000. Ensemble methods in machine learning. *Multiple Classifier Systems.* Berlin: Springer-Verlag Berlin.

EFRON, B. & TIBSHIRANI, R. J. 1993. *An introduction to the Bootstrap,* New York, USA, Chapman & Hall.

FLEISS, J. L., : 1981. Statistical Methods for Rates and Proportions.John Wiley & Sons.

FRANCIA, S. S. & GARCÍA, M. N. M. Marzo, 2006. Multiclasificadores: Métodos y Arquitecturas.Departamento de Informática y Automática Universidad de Salamanca.

FREUND, Y. & SCHAPIRE, R. E. 1997. Decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences,* 55**,** 119–139.

GIACINTO, G. & ROLI, F. 2001. Design of effective neural network ensembles for image classification purposes.

. *Image and Vision Computing, 19(9-10):* **,** p. 699-707.

GONZÁLEZ, E. 2010. Proyección del Centro de Desarrollo de la Electrónica hacia la Comunidad (PRODEC).Universidad Central de Las Villas.

HANSEN, L. K. & SALAMON., P. 1990. . Neural Network Ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence***,** 12: p. 993-1001.

HO, T. K. 1998. The random space method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 20**,** 832-844.

KOHAVI, R. & WOLPERT., D. H. 1996. Bias Plus Variance Decomposition for Zero-One Loss Functions in Machine Learning. *Proceedings of the Thirteenth International Conference.* .

KUNCHEVA, L. I. 2004. *Combining Pattern Classifiers, Methods and Algorithms,* New York, NY, Wiley Interscience.

KUNCHEVA, L. I. & WHITAKER, C. J. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning,* 51**,** 181-207.

MITCHELL, T. M. 1997. *Machine Learning*, McGraw-Hill Science/Engineering/Math; (March 1, 1997).

PARTRIDGE, D. & KRZANOWSKI, W. 1997. Software diversity: practical statistics for its measurement and exploitation. *Information and Software Technology***,** 39(10): p. 707-717.

PARTRIDGE, D. & KRZANOWSKI, W. 1997. "Distinct Failure Diversity in Multiversion Software".

POLIKAR, R. 2006. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine.* 6(3) ed.

QUINLAN, J. R. 1986. Induction of decision trees. *Machine Learning,* 1**,** 81-106.

QUINLAN, J. R. 1993. *C4.5: Programs for Machine Learning.,* San Mateo,CA, Morgan Kaufmann.

RUTA, D. & GABRYS, B. 2001. Analysis of the Correlation Between Majority Voting Error and the Diversity

Measures in Multiple Classifier Systems, in Soft Computing and Intelligent Systems for Industry: Proceedings

and Scientific Program:. *Fourth International ICSC Symposium 20012001 ICSC-NAISO Academic Press:*

Paisley, Scotland. p. 50.

SALAZAR, S. 2005. *NEngine v1.0. Una Herramienta Software para Redes Neuronales Recurrentes.* Universidad Central "Marta Abreu" de Las Villas.

SCHAPIRE, R. E. 1990. The strength of weak learnability. *Machine Learning,* 5**,** 197-227.

SKALAK, D. B. 1996. The Sources of Increased Accuracy for Two Proposed Boosting Algorithms, .

SKURICHINA, M. 2001. *Stabilizing weak classifiers.* PhD, Delft University of Technology.

VAPNIK, V. 1995. *The Nature of Statistical Learning Theory,* New York, Springer-Verlag.

WITTEN, I. & FRANK, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques,* San Francisco, Diane Cerra.

WOLPERT, D. 1992. Stacked generalization. *Neural Networks,* 5**,** 241-259.