UNA REVISIÓN DE LOS MÉTODOS DE CLASIFICACIÓN APLICABLES A LA ECONOMÍA.

Esteban Alfaro Cortés Matías Gámez Martínez Noelia García Rubio
Profesor Asociado Prof. Titular de Universidad Profesora Asociada
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 recogen algunos de los métodos de clasificación que más se utilizan, exponiendo brevemente en qué consiste cada uno de ellos y comparándolos en función de los resultados que obtienen. Entre los métodos paramétricos destaca el análisis discriminante que es la técnica estadística clásica para la clasificación. En él se establece una superficie que separa los ejemplos de las distintas clases. De los métodos no paramétricos estudiamos en primer lugar la clasificación por el vecino más próximo, este método es el más sencillo y clasifica cada observación asignándole la misma clase de sus vecinos. Las redes neuronales artificiales son un método más sofisticado aunque muy efectivo, e intentan en cierto modo imitar el funcionamiento del sistema nervioso humano, mediante unidades simples interconectadas formando una red donde el aprendizaje se realiza en paralelo. Por último, los árboles de clasificación realizan una partición recursiva del espacio de ejemplos a través de los nodos, llevándolos a lo largo de las ramas que nacen en ellos hasta alcanzar un nodo hoja donde sea posible asignar una clase.

Palabras clave: Métodos de clasificación, Análisis Discriminante, Vecino más próximo, Redes Neuronales, Árboles de clasificación.

1. INTRODUCCIÓN.

El objetivo de este artículo es proporcionar una revisión de los métodos de clasificación que más se utilizan actualmente, comentando tanto aspectos generales de la clasificación, como cada uno de los métodos de forma algo más detallada, recogiendo y analizando por último, algunas de las comparaciones más interesantes que se han hecho de los mismos.

El problema de la clasificación está presente en un amplio rango de la actividad humana. En su forma más general el término puede cubrir cualquier contexto en el que se toma una decisión o se realiza una predicción en base a la información disponible en ese momento,

y un procedimiento de clasificación es entonces un método formal para repetir esos razonamientos en situaciones nuevas.

En este trabajo, nos centramos en una interpretación algo más concreta. El problema consiste en construir un procedimiento que se aplicará sobre una secuencia conjunta de casos en los que cada nuevo caso debe ser asignado a una de entre un conjunto de clases predefinidas, basándose en atributos o características observadas.

La construcción de un sistema de clasificación a partir de un conjunto de datos para el cual se conocen las verdaderas clases ha sido denominada de diferentes formas: reconocimiento de patrones, análisis discriminante o aprendizaje supervisado.

El problema de clasificación se presenta en contextos tan distintos como, por ejemplo, el procedimiento mecánico para enviar las cartas en base a la lectura automática de los códigos postales, la toma de decisiones respecto a las solicitudes de crédito de los individuos en base a información financiera y otra información personal o el diagnóstico preliminar de la enfermedad de un paciente para seleccionar un tratamiento inmediato mientras se espera a los resultados definitivos de las pruebas.

2. DEFINICIÓN DE CLASIFICACIÓN.

Cuando hablamos de clasificación puede tener dos significados distintos. Bien podemos tener un conjunto de observaciones con el objetivo de establecer la existencia de clases o grupos en los datos. O bien podemos saber que existen determinadas clases y que el objetivo sea establecer una regla por la que podamos clasificar una nueva observación dentro de una de las clases existentes. El primer tipo se conoce como *aprendizaje no supervisado* (o *clustering*), el segundo como *aprendizaje supervisado*. Nosotros nos centraremos en el segundo, el aprendizaje supervisado. En Estadística, tradicionalmente se ha utilizado para este propósito el análisis discriminante pero en los últimos tiempos se han desarrollado nuevas técnicas, en parte gracias al desarrollo experimentado por la capacidad de los soportes informáticos.

Los métodos de clasificación han sido utilizados en muchas disciplinas, por ejemplo,

en biología (genética), medicina (diagnóstico de enfermedades), astronomía, ingeniería, control y robótica. En economía son muchas las posibles aplicaciones de los métodos de clasificación, que intentan comprender la pertenencia a un grupo, por ejemplo, separación de los individuos en clientes y no clientes para una empresa, clasificación de las empresas en rentables y no rentables, determinación de las claves que conducen a un nuevo producto al éxito o al fracaso. Establecer la categoría de riesgo asociada a una solicitud de crédito, tanto en el caso de créditos personales como de empresas. Los inversores deben decidir comprar o vender las acciones en base a la información de la propia empresa y de la economía en general. Las autoridades fiscales deben decidir cuándo hay que realizar una inspección basándose en información financiera y fiscal tanto en el caso de empresas como de personas. Las autoridades financieras (Banco de España, Comisión Nacional del Mercado de Valores, etc.) tienen que decidir sobre la intervención o no de una institución financiera. En el mercado de la vivienda, se puede establecer un sistema de clasificación que determine de forma objetiva el grado de calidad de una vivienda. En el marco de la economía regional es interesante establecer los parámetros que determinan la pertenencia de una región a un grupo u otro de regiones, ya que en función de esto se determinan las ayudas a recibir. En control de calidad se distingue entre productos defectuosos y no defectuosos.

3. CLASIFICACIÓN DE LOS PROCEDIMIENTOS DE CLASIFICACIÓN.

Algunos de los procedimientos de clasificación (discriminante lineal, árboles de decisión y basados en reglas, vecino más próximo) fueron desarrollados hace ya algunos años. Como es lógico, han sido refinados y extendidos, pero actualmente todavía representan las ramas más importantes en clasificación, tanto a nivel aplicado como a nivel de investigación. Los procedimientos que citamos a continuación pueden relacionarse directamente a uno u otro de los anteriores. Sin embargo, tradicionalmente se han dividido en cuatro grupos, estadística clásica, técnicas estadísticas modernas, aprendizaje automático y redes neuronales. Para algunos de los métodos la inclusión en uno u otro grupo resulta un poco arbitraria.

3.1. El discriminante lineal y sus extensiones.

El discriminante lineal es uno de los procedimientos de clasificación más antiguos y es el que mayoritariamente se utiliza en los paquetes estadísticos. La idea es dividir el espacio muestral mediante una serie de lineas en el espacio bidimensional, planos en el caso de tres dimensiones y, en general, hiperplanos para muchas dimensiones. La línea que divide dos clases se define, de tal forma que corte por la mitad la línea que une los centros de esas clases, la dirección de la línea se determina por la forma de los grupos de puntos. La recta, el plano o hiperplano será una combinación lineal de las variables que caracterizan a los ejemplos. Al ser un procedimiento geométrico, la naturaleza de los datos es importante, debiendo ser variables cuantitativas.

Podemos incluir en este grupo aquellos procedimientos que empiezan con combinaciones lineales de las medidas, incluso si estas combinaciones están sujetas después a transformaciones no lineales. Podemos citar seis procedimientos más de este tipo: discriminante logístico, discriminante cuadrático, perceptrón multicapa (backpropagation y cascade), DIPOL92 y projection pursuit. Nótese que este grupo contiene métodos estadísticos y redes neuronales (concretamente perceptrón multicapa)

3.2. Árboles de decisión y métodos basados en reglas.

Los árboles de clasificación están compuestos por nodos y ramas. Cada nodo representa una cuestión o decisión sobre una de las características de los ejemplos. El nodo inicial se suele llamar nodo raíz. De cada nodo pueden salir dos o más ramas dependiendo de que la respuesta a la cuestión planteada sea binaria o no. Finalmente, se alcanzan los nodos terminales u hojas y se toma una decisión sobre la clase a asignar. El desarrollo del árbol se realiza de modo que la concentración en los nodos (calculada a través del índice de Gini o de la Entropía) vaya aumentado. La naturaleza de la información es poco relevante, siendo un procedimiento ideal para datos cualitativos.

En este grupo de pueden incluir gran cantidad de procedimientos entre los que podemos destacar: NewID, AC^2, Cal5, CN2, C4.5, CART, IndCART, árbol de Bayes e Itrule.

3.3. Estimaciones de la densidad.

Este grupo es un poco menos homogéneo, pero sus miembros tienen en común que el procedimiento está muy relacionado con la estimación de la densidad local de probabilidad en cada punto del espacio muestral. Podemos citar: vecino más próximo, funciones de base radial, Bayes simplificado, árboles múltiples, mapas autoorganizados de Kohonen, aprendizaje por vectores de cuantificación (LVQ). Relacionado con los anteriores, el método de estimación de la densidad del núcleo, aunque a diferencia de ellos, aquí se estima la densidad de probabilidad para un núcleo. Este grupo también contiene sólo métodos estadísticos y redes neuronales.

Por su importancia debemos explicar brevemente en qué consisten el método del vecino más próximo y las redes neuronales. El primero se basa en la idea de que lo más probable es que las observaciones de la misma clase estén cercanas entre sí. Para una nueva observación, podemos fijarnos, por ejemplo, en las *k* observaciones más cercanas de todas las almacenadas previamente y clasificaremos la nueva observación de acuerdo con la clase mayoritaria entre sus vecinos.

Las Redes Neuronales Artificiales (RNA) son un método no paramétrico de clasificación, que permite obtener un grado elevado de precisión. Esta técnica pretende imitar el funcionamiento del cerebro humano, en lo referente a su estructura neuronal, que de manera simplificada constituye una red de neuronas interconectadas lo que posibilita el procesamiento en paralelo. Así, en el caso de las redes neuronales artificiales los nodos, o unidades equivalentes a las neuronas, reciben como entradas la suma ponderada de las salidas de otras unidades, que es como se cree que actúan las neuronas humanas.

4. COMPARACIÓN EMPÍRICA DE LOS MÉTODOS DE CLASIFICACIÓN.

Hemos estudiado los métodos de aprendizaje más utilizados, pero hay otras técnicas de aprendizaje que podríamos citar, y muchas nuevas están en el proceso de ser formuladas o descubiertas. Para medir el comportamiento generalmente se utiliza la tasa de error, de esta forma podemos comparar de forma objetiva los distintos métodos, aunque existen otras características importantes en un sistema de clasificación además de la precisión, como por ejemplo su comprensibilidad, el tiempo necesario para entrenarlo o el tiempo que emplea

para clasificar nuevas observaciones una vez entrenado.

Son muchas las comparaciones que se han realizado, pero existen algunos aspectos con los que debemos tener cuidado a la hora de analizar y tener en cuenta las mismas. En algunas de ellas, la elección de los algoritmos a comparar deja fuera a alguno de los que hemos citado en este trabajo, siendo esto una desventaja importante. En ese caso se pierde la generalidad y debe entenderse como una comparación entre determinados métodos de forma concreta.

En muchos casos, los autores han desarrollado su propio algoritmo y son expertos en su campo, pero no lo son tanto en otros métodos, lo que puede producir un sesgo natural en contra de los otros métodos.

Los algoritmos elegidos pueden no representar los últimos avances. Debido al elevado ritmo con el que se descubren y desarrollan nuevas técnicas en todas esas áreas podemos estar comparando la red neuronal más avanzada con un árbol de clasificación pasado de moda como el sistema ID3 (superado por la versión posterior C4.5).

Los conjuntos de datos disponibles son normalmente pequeños o simulados y por tanto no muy representativos de las aplicaciones de la vida real. La elección de los conjuntos de datos, especialmente en el caso de datos simulados, proporciona un sesgo a favor de ciertos algoritmos.

También a menudo la elección de los criterios para la comparación está sesgada en favor de un tipo de algoritmo, a veces incluso utilizando criterios de coste poco realistas.

Otras veces puede haber problemas debido a las diferencias en la forma en que los datos son preprocesados, por ejemplo, eliminando o reemplazando los valores omitidos, o al transformar los atributos categóricos en variables numéricas. Las definiciones de las clases pueden ser más adecuadas para algunos algoritmos que para otros. Además, las proporciones de las clases en el conjunto de entrenamiento pueden ser sustancialmente diferentes de los valores poblacionales (a veces incluso de forma deliberada). Algunos de estos estudios comparativos utilizan variaciones de los conjuntos de datos y algoritmos originales.

En la medida de lo posible sería deseable evitar todos estos problemas a la hora de realizar una comparación entre los diferentes métodos de clasificación, de todas formas la presencia de alguno de estos problemas no anula el valor de la comparación hecha, simplemente debe relativizar las conclusiones extraídas a partir del mismo.

4.1. Algunas comparaciones.

A continuación recogemos algunas de las comparaciones más interesantes de las que se han hecho tanto como por su contenido, métodos de clasificación y conjuntos de datos que utiliza, como por el análisis realizado y las conclusiones que obtienen. En primer lugar, por orden cronológico, está el trabajo de Weiss y Kulikowsky (1991) que en la línea del trabajo de Weiss con Kapouleas en 1989, recoge una comparación empírica bastante interesante. En segundo lugar, recogemos lo aportado por Michie, Spiegelhalter y Taylor en 1994, que desde la perspectiva del aprendizaje automático analizan los resultados del proyecto Statlog. Por último, Lim, Loh y Shih en 1998 analizan de forma más amplia el comportamiento de treinta y tres clasificadores en dieciséis conjuntos, teniendo en cuenta también el tiempo de computación además de la tasa de error de los mismos.

4.1.1. Sistemas automáticos que aprenden.

Una comparación muy interesante es la que proporciona Weiss (1991) que utiliza una muestra representativa de los distintos tipos de problemas que más frecuentemente podemos encontrarnos al hablar de clasificación. Los problemas que utiliza son de dimensión relativamente baja, aunque realistas, cada problema tiene relativamente pocas variables y clases. Como es habitual las aplicaciones están caracterizadas por incertidumbre en la clasificación. En algunos casos, como el problema del cáncer, las características son relativamente débiles y realizar buenas predicciones se hace poco probable. En otros, como la aplicación del conjunto de datos de tiroides, las características son bastante fuertes y casi es posible realizar predicciones libres de error (cien por cien seguras).

Para los conjuntos de datos más pequeños, utilizan remuestreo para estimar la tasa de error. Alrededor de los cien casos, las técnicas de remuestreo, como la validación cruzada, puede obtener buenas estimaciones de la tasa de error real. De hecho, los datos del estudio de los lirios han sido utilizados durante muchos años y las comparaciones se han realizado

sobre la base del error estimado dejando uno fuera.

En esos experimentos, como grupo los métodos de árboles de clasificación lo hicieron mejor y las redes neuronales de retropropagación tuvieron un comportamiento ligeramente peor. Para las aplicaciones sencillas, una solución basada en un conjunto de reglas fue mejor que el resto, aunque para problemas más complejos los árboles de decisión obtienen mejores resultados. El problema más grande estudiado, la aplicación del tiroides, está algo sesgado a favor de las soluciones basadas en reglas. Los diagnósticos se derivan de un sistema basado en reglas que al parecer utiliza los mismos umbrales de las pruebas de laboratorio para especificar altos o bajos niveles para todas las hipótesis.

En los experimentos de Weiss y Kulikowsky, los clasificadores estadísticos se comportan de forma consistente a lo que se espera de ellos. El clasificador lineal normal, obtiene buenos resultados en todos los casos salvo en el experimento del tiroides. Estos clasificadores son ampliamente utilizados por su sencillez y porque el porcentaje de error de entrenamiento normalmente se mantiene bien en los casos de prueba. La extensión natural, el discriminante cuadrático, se ajusta mejor para datos que se distribuyen según la normal, pero empeora rápidamente para datos que no siguen esta distribución. Se comporta pobremente en la mayoría de esos experimentos. El método del vecino más próximo lo hace bien con buenos atributos pero tiende a empeorar con muchos atributos malos.

Las redes neuronales se comportan bien y al contrario que los clasificadores estadísticos obtienen buenos resultados en el problema del tiroides. Sin embargo, después de todo no logran los mejores resultados y consumen cantidades enormes de tiempo de ordenador especialmente cuando el algoritmo de retropropagación se aplica en la manera estándar. Las soluciones de las redes son a veces igualadas en su comportamiento por clasificadores sencillos.

Todos los métodos que emplean una medida de ajuste de complejidad siguen el patrón clásico para los sistemas de aprendizaje. Por ejemplo, en las redes neuronales, puede verse una clara relación entre el número de unidades ocultas y las tasas de error aparente y estimada. Conforme aumenta el número de unidades ocultas la tasa de error aparente disminuye. Sin embargo, en algún punto, cuando el clasificador sobreajusta los datos, la curva de la verdadera tasa de error se mantiene constante e incluso comienza a aumentar.

El mismo comportamiento puede observarse en los árboles de clasificación cuando aumenta el número de nodos.

La siguiente tabla muestra los resultados de la tasa de error aparente y la estimada mediante validación cruzada con 10 particiones, para los clasificadores en los cuatro conjuntos citados:

	Lirios		Apendicitis		Cáncer		Tiroides	
	Apar	V.C.	Apar	V.C.	Apar	V.C.	Apar	V.C.
D.L.	0,020	0,020	0,113	0,132	0,254	0,294	0,0615	0,0615
D.C.	0,020	0,027	0,217	0,264	0,245	0,344	0,1031	0,1161
NN	0,000	0,040	0,000	0,179	0,000	0,347	0,0000	0,0423
BP	0,017	0,033	0,098	0,142	0,243	0,285	0,0050	0,0146
CART	0,040	0,047	0,094	0,151	0,226	0,229	0,0021	0,0064

Tabla 1. Resultados de la tasa de error aparente y la estimación mediante validación cruzada

4.1.2. Aprendizaje automático, clasificación neuronal y estadística.

Michie, Spiegelhalter y Taylor en 1994 consideran cuatro grupos de aplicaciones, según ellos porque eso facilita la descripción y la interpretación y además porque los resultados de algunos clasificadores se ven alterados por el tipo de aplicación. Los cuatro grupos son: conjuntos con costes, conjuntos de riesgo crediticio, conjuntos de imágenes y otros.

En este trabajo nos hemos centrado en los métodos de clasificación que no tienen en cuenta el coste asociado a cada clase o a los errores en la clasificación. Además de todos los sistemas de clasificación que utilizan estos autores nosotros nos limitaremos a aquellos expuestos en este trabajo sin que ello suponga un menosprecio hacia los demás.

Conjuntos de crédito.

En estos problemas el objetivo es predecir el riesgo asociado a las solicitudes de crédito. Son dos problemas, gestión de créditos del Reino Unido y riesgo crediticio de Australia (Quinlan 1993)

	C. Aust	C.Gestión
D. Log	0,141	0,030
C4.5	0,155	0,022
B.P.	0,154	0,023
D. Lineal	0,141	0,033
k-NN	0,181	0,031
CART	0,145	-
D. Cuad	0,207	0,05
R. Defec	0,444	0,05

Tabla 2. Tasas de error en los conjuntos de crédito.

En el análisis completo tres de los seis mejores algoritmos son árboles de clasificación (C4.5 es uno de ellos) y el segundo lugar lo ocupa una variante de las redes neuronales. Se puede concluir que los árboles de clasificación actúan razonablemente bien en conjuntos de créditos, aunque no se dispone del resultado del algoritmo CART en el conjunto de gestión de créditos. Una posible explicación para la buena actuación de los árboles de clasificación, es que fue un humano el que clasificó los datos del conjunto de entrenamiento y en muchas ocasiones el proceso de decisión humano realiza una partición recursiva del espacio en función de una serie de atributos como hacen también los árboles clasificación. Además los árboles de clasificación se comportan bien cuando hay muchos atributos binarios o categóricos.

Reconocimiento de imágenes.

Los problemas de reconocimiento de imágenes son aplicables en una amplia variedad de contextos. En algunos casos hay que clasificar la imagen entera (o un objeto en la imagen), mientras que en otros la clasificación se realiza a partir de la segmentación de una imagen. Michie et al. consideran por separado ambos grupos. En el primero, el reconocimiento de objetos, incluyen cinco conjuntos, en concreto el conjunto de Karhunen-Loeve, el de los dígitos, la silueta de vehículos, los cromosomas y la escritura manual.

	KL	Dígitos	Vehic	Crom	Letras
D. Cuad	0,025	0,054	0,15	0,084	0,113
k-NN	0,020	0,047	0,275	0,123	0,070
D. Log	0,051	0,051	0,192	0,131	0,234
D. Lineal	0,075	0,075	0,216	0,107	0,302
B.P.	0,049	0,049	0,207	-	0,327
C4.5	0,180	0,18	0,266	0,175	0,132
CART	-	0,16	0,235	-	-
R. Defec	0,900	0,9	0,75	0,960	0,960

Tabla 3. Tasas de error en los conjuntos de reconocimiento de objetos.

Estos autores señalan su decepción por el comportamiento del algoritmo de retropropagación en estos problemas, ya que los que propusieron este método destacaron su habilidad para modelar comportamientos no lineales. Algunos de estos conjuntos son significativamente no lineales y aunque, es cierto que este algoritmo lo hace bien, otros algoritmos más sencillos presentan un comportamiento mejor. Considerando el esfuerzo requerido para optimizar el entrenamiento del algoritmo de retropropagación, se preguntan cuando supone realmente una ventaja sobre métodos tradicionales.

El discriminante cuadrático y los k vecinos más próximos son los métodos que en general se comportan mejor. Destaca el pobre comportamiento de los árboles de clasificación en este caso.

Reconocimiento de imágenes segmentadas.

En este grupo consideran cuatro problemas: imágenes de satélite (satim), segmentos de imágenes al aire libre (segm) y dos problemas de separación de los caracteres que forman una palabra (Cut50 y Cut20)[1].

La siguiente tabla proporciona las tasas de error para los cuatro problemas de segmentación. Los árboles de clasificación lo hacen muy bien en estos conjuntos y los métodos estadísticos tradicionales obtienen resultados bastante modestos. Una probable

[1] El conjunto de datos Cut20 consiste en los primeros veinte atributos del conjunto Cut50 ordenados según su importancia en un procedimiento de regresión secuencial (stepwise).

explicación es que estos conjuntos de datos se originan como problemas de partición lo que facilita la labor de los árboles de clasificación.

	Satim	Segm	Cut20	Cut50
k-NN	0,094	0,077	0,036	0,027
C4.5	0,150	0,040	0,036	0,035
B.P.	0,139	0,054	0,043	0,041
D. Log	0,163	0,109	0,046	0,037
CART	0,138	0,040	-	-
D. Lineal	0,171	0,116	0,050	0,050
D. Cuad	0,155	0,157	0,088	0,097
R. Defec	0,760	0,857	0,060	0,060

Tabla 4. Tasas de error en los conjuntos de segmentación de imágenes.

Los autores resaltan la comparación entre los resultados de Cut20 y Cut50, ya que es de esperar que el comportamiento de un clasificador empeore cuando el número de atributos disminuye, puesto que disminuye la información contenida. El discriminante cuadrático obtiene un mal resultado en el conjunto Cut20 pero aún peor en el Cut50. Los árboles de clasificación se comportan mejor en el conjunto Cut50, y utilizan árboles de clasificación más pequeños para obtener mayor precisión. Hay que destacar el buen comportamiento que el método del vecino más próximo de orden k tiene en los conjuntos de imágenes.

Otros conjuntos.

	Belg	NewBel	Tset	Diab	ADN	Fallos	Shutt (%)	Tech
C4.5	0,040	0,018	0,049	0,270	0,076	0,305	0,100	0,120
D. Log	0,007	0,028	0,117	0,223	0,061	0,221	3,830	0,401
CART	0,034	0,022	0,041	0,255	0,085	0,318	0,080	-
B.P.	0,017	0,022	0,065	0,248	0,088	0,228	0,430	-
D. Lineal	0,025	0,041	0,122	0,225	0,059	0,204	4,830	0,391
D. Cuad	0,052	0,035	0,098	0,262	0,059	0,293	6,720	0,495
k-NN	0,059	0,052	0,057	0,324	0,155	0,375	0,440	0,204
R. Defec	0,362	0,074	0,490	0,350	0,480	0,610	21,40	0,770

Tabla 5. Tasas de error en el resto de conjuntos.

Dos de estos problemas son problemas de partición puros (Shuttle y Technical) con fronteras paralelas a los ejes de los atributos. Otros dos son conjuntos simulados (Belgian y Belgian power II) y se pueden describir como algo intermedio entre predicción y partición. El objetivo del conjunto de datos Tsetse es dividir un mapa en dos regiones y repetir una división dada de forma tan precisa como se pueda. Además, este conjunto tiene algo de artificial ya que algunos atributos han sido construidos por interpolación a partir de una pequeña cantidad de información.

El problema de diabetes es un problema de predicción. Los otros dos son los conjuntos del ADN y de los fallos de máquinas que son una mezcla entre división, predicción y discriminación.

Aunque los propios autores señalan que es arriesgado extraer conclusiones de un grupo tan heterogéneo de problemas, resaltan que los árboles de clasificación son los que mejor se comportan seguidos por el algoritmo de retropropagación. Mientras que las técnicas estadísticas clásicas obtienen resultados discretos en general, siendo el mejor el discriminante logístico, el método del vecino más próximo es en este caso el método peor valorado.

En términos generales Michie et al. concluyen que las redes neuronales tiene buenos resultados cuando los métodos estadísticos también los tienen y vice versa, es decir, tienden a comportarse bien en la misma clase de conjuntos. Señalando que las redes neuronales son más parecidas a los procedimientos estadísticos que a los árboles de clasificación lo cual es coherente con la clasificación propuesta en el apartado 3 de este trabajo.

4.1.3. Una comparación empírica de árboles de clasificación y otros métodos de clasificación.

En este trabajo Lim, Loh y Shih comparan 22 árboles de clasificación, 9 técnicas estadísticas y dos redes neuronales en dieciséis conjuntos de datos en términos del error de clasificación, del tiempo de computación y en el caso de los árboles el número de nodos terminales. Estos autores concluyen que las medias de las tasas de error para la mayoría de los clasificadores no tienen diferencias estadísticamente significativas mientras que en tiempo de computación las diferencias son mucho más amplias.

Nosotros nos centraremos en los resultados para los clasificadores que hemos visto, pero teniendo en cuenta algunas matizaciones. El sistema CART de árboles de clasificación lo utilizan de dos formas según se aplique o no la regla 1-SE de podado[2], y lo representan por IC1 e IC0, respectivamente. En cuanto a las redes neuronales utilizan dos sistemas alternativos al de retropropagación, como son el LVQ (Learning Vector Quantization) y el RBF (Radial Basis Function) y aunque no los hemos visto en este trabajo los recogemos para tener una idea aproximada del comportamiento de las redes neuronales.

Vamos a describir brevemente cada uno de los conjuntos que utilizan estos autores:

bcw : el conjunto de datos del cáncer de mama fue recogido en la Universidad de Wisconsin. Hay que predecir a partir de una muestra de tejido del pecho de una paciente cuándo el tumor es maligno o benigno.

cmc: elección del método anticonceptivo. El problema es predecir el método anticonceptivo elegido por una mujer en base a sus características demográficas o socioeconómicas. Hay tres clases: no utiliza, utiliza métodos a largo plazo y utiliza métodos a corto plazo.

dna: en este problema se trata de reconocer las fronteras entre la parte de ADN que permanece durante el proceso de creación de proteínas y la parte de ADN que cambia.

hna: se pretende aprender a diagnosticar si un paciente padece o no una enfermedad cardiaca en función de varias pruebas médicas realizadas al mismo.

bos: precio de la vivienda en Boston. Describe el precio de las viviendas en los distritos de Boston como función de doce atributos continuos y uno categórico. La clase de cada ejemplo es el valor medio de la vivienda segmentado en tres intervalos.

led: hay que predecir cuál de los dígitos decimales (del 0 al 9) es el que forma un conjunto de siete diodos luminosos que son representados por siete atributos dicotómicos, que toman el valor 1 si el diodo está encendido y 0 en caso contrario.

[2] La regla 1-SE es una técnica de podado que selecciona el árbol más pequeño de entre los que tienen una precisión superior a la del mejor disminuida en una desviación estándar.

bld: en este problema se trata de predecir la cantidad de alcohol que ha ingerido un varón, medida en términos de medias pintas. Para determinarlo se emplean variables relacionadas con las pruebas sanguíneas a las que se somete al paciente.

pid: se trata de predecir cuándo un paciente dará positivo en las pruebas de diabetes de acuerdo a la Organización Mundial de la Salud a partir de las medidas fisiológicas y determinadas pruebas médicas.

sat: imágenes de satélite de Statlog. Se trata de predecir la clasificación de un píxel en función de la clase de sus vecinos en un espectro de cuatro bandas.

seg: imágenes segmentadas de Statlog. Las observaciones se obtienen de forma aleatoria a partir de una base de datos de siete imágenes del exterior. Hay que clasificar cada píxel diciendo de qué tipo de imágenes se trata.

smo: actitud hacia las restricciones a fumar. El problema es predecir la actitud hacia las restricciones sobre el tabaco en el lugar de trabajo a partir de ciertas características de los individuos.

thy: hay que diagnosticar enfermedades del tiroides, en concreto hay tres clases, normal, hiperfuncionamiento y funcionamiento por debajo de lo normal.

veh: silueta de vehículos de Statlog. El problema es clasificar una silueta dada en uno de los cuatro tipos de vehículos, utilizando un conjunto de características extraídas de la silueta del vehículo.

vot: votos recogidos en el congreso de los Estados Unidos en 1984. Se pretende descubrir el signo político de los congresistas (republicano o demócrata) a partir de la votación en dieciséis temas clave.

wav: forma de la onda. Este es un problema artificial. Consta de tres clases y está basado en tres formas distintas de ondas.

tae: los datos consisten en evaluaciones del comportamiento de la enseñanza durante tres semestres regulares y dos semestres de verano de 151 tareas auxiliares de la enseñanza en el departamento de estadística de la universidad de Wisconsin-Madison.

La siguiente tabla resume la información más relevante sobre las características de los conjuntos.

Conj de datos	Nº de casos	Nº de clases	Número de atributos							
			Continuos	Categóricos						Total
				2	3	4	5	25	26	
bcw	683	2	9							9
cmc	1473	3	2	3		4				9
dna	2000	3				60				60
hea	270	2	7	3	2	1				13
bos	506	3	12	1						13
led	2000	10		7						7
bld	345	2	6							6
pid	532	2	7							7
sat	4435	6	36							36
seg	2310	7	19							19
smo	1855	3	3	3	1		1			8
thy	3772	3	6	15						21
veh	846	4	18							18
vot	435	2			16					16
wav	600	3	21							21
tae	151	3	1	2				1	1	5

Tabla 6. Descripción de los conjuntos de datos utilizados.

En la tabla 7 se recoge la tasa de error estimada en cada conjunto para los distintos clasificadores.

	C4.5	IC0	IC1	LDA	QDA	NN	LOG	LVQ	RBF	Defecto
bcw	0,043	0,045	0,047	0,039	0,048	0,048	0,034	0,028	0,034	0,350
cmc	0,483	0,451	0,449	0,492	0,541	0,601	0,489	0,491	0,458	0,573
dna	0,076	0,058	0,062	0,060	0,137	0,317	0,064	0,132	0,293	0,492
hea	0,196	0,207	0,219	0,141	0,248	0,226	0,159	0,341	0,193	0,444
bos	0,221	0,254	0,266	0,249	0,266	0,227	0,243	0,314	0,225	0,657
led	0,271	0,279	0,286	0,271	0,273	0,294	0,269	0,313	0,446	0,890
bld	0,308	0,327	0,319	0,326	0,401	0,370	0,309	0,329	0,330	0,419
pid	0,242	0,237	0,239	0,221	0,238	0,295	0,230	0,243	0,230	0,333
sat	0,146	0,138	0,154	0,160	0,141	0,217	0,163	0,098	0,121	0,765
seg	0,032	0,037	0,532	0,083	0,123	0,022	0,043	0,084	0,126	0,857
smo	0,305	0,319	0,305	0,305	0,454	0,410	0,305	0,366	0,307	0,305
thy	0,006	0,007	0,006	0,062	0,890	0,065	0,041	0,071	0,032	0,073
veh	0,277	0,265	0,298	0,224	0,145	0,224	0,196	0,374	0,372	0,739
vot	0,048	0,048	0,044	0,046	0,055	0,053	0,050	0,050	0,052	0,386
wav	0,261	0,297	0,313	0,178	0,179	0,396	0,154	0,170	0,151	0,667
tae	0,503	0,451	0,537	0,411	0,543	0,349	0,450	0,628	0,464	0,656

Tabla 7. Tasas de error estimadas.

Para estimar la tasa de error de los clasificadores en aquellos conjuntos suficientemente grandes (tamaño muy superior a mil y conjunto de prueba de al menos mil) se utiliza un conjunto de prueba. Es decir, el clasificador se construye utilizando las observaciones en el conjunto de entrenamiento y después se evalúa en el conjunto de prueba. Seis de los dieciséis conjuntos se evaluaron de esta forma. En el resto, se utiliza validación cruzada con diez particiones para estimar el error.

Lim, Loh y Shih proporcionan una ordenación de los distintos clasificadores en función de la tasa de error media a lo largo de todos los conjuntos. Esta ordenación, aunque interpretada con cierta cautela pues no tiene en cuenta ni el tamaño ni la dificultad de los conjuntos para ponderar la media, sí puede servir a título orientativo.

D.Log (0,204), D.Lin.(0,208), IC0 (0,215), C4.5(0,220), IC1(0,227), RBF(0,257), LVQ (0,269), NN(0,281), D.Cuad (0,301).

Según esta ordenación, las redes neuronales presentan en general un comportamiento modesto al igual que el método del vecino más próximo, lo que en principio era previsible dada su simplicidad. Por el contrario los árboles de clasificación y en especial las técnicas estadísticas tradicionales como el discriminante logístico y lineal son los que mejores resultados obtienen. De manera más particular los árboles de clasificación (C4.5 y CART) obtienen el mejor resultado en un total de 8 conjuntos (cada uno en 4). Las técnicas estadísticas tradicionales logran la menor tasa de error en 7 conjuntos (3 el discriminante lineal, 1 el discriminante cuadrático y 2 el discriminante logístico). Los sistemas basados en redes neuronales sólo consiguen vencer en tres conjuntos (LVQ en 2 y RBF en 1). Por último el vecino más próximo de orden k obtiene la mayor precisión en dos conjuntos.

Hay que destacar el conjunto que recoge la actitud hacia las restricciones a fumar ya que en él se produce un quíntuple empate entre discriminante lineal, el discriminante logístico, el sistema de árboles de clasificación C4.5 y el sistema CART aplicando la regla 1-SE para el podado con la regla por defecto que como hemos visto consiste en asignar en todos los casos la clase mayoritaria en el conjunto de entrenamiento. Podemos afirmar que ese problema es el más difícil para los clasificadores ya que ninguno consigue superar la precisión de la regla por defecto.

5. CONCLUSIONES.

Para acabar, podemos destacar los siguientes aspectos de este trabajo.

En primer lugar, los métodos de clasificación tienen múltiples aplicaciones en muy diversos campos, tales como la biología, la medicina, astronomía, ingeniería, control y robótica. Pero pueden ser especialmente útiles en algunas áreas de la economía, como la economía regional, determinadas decisiones financieras y fiscales, para establecer tipologías de empresas o clientes, para fijar objetivamente la calidad de las viviendas o de los productos fabricados.

En segundo lugar, hemos mostrado que existen varias alternativas a la hora de elegir el método de clasificación que se va a utilizar, desde los métodos paramétricos como el discriminante lineal, a otros métodos no paramétricos como el método del vecino más próximo, las redes neuronales artificiales o los árboles de clasificación. Para ello es también fundamental conocer la naturaleza de los datos.

Por último, las comparaciones empíricas realizadas no se ponen de acuerdo en la supremacía de un método sobre el resto, sino que a lo sumo establecen determinados tipos de problemas para los que un método se comporta mejor que los otros. Aconsejando ante un nuevo problema la aplicación de los distintos métodos disponibles y elegir el que mejor se comporte para ese problema en particular.

6. BIBLIOGRAFÍA.

BISHOP, C. M. (1995), *Neural Networks for Pattern Recognition*. New York: Oxford University Press Inc.

BREIMAN, L.; FRIEDMAN, J.H.; OLSHEN, R. Y STONE, C.J. (1984), *Clasification and regresion trees.* Belmont, Wadsworth International Group.

DEVIJVER, P.A. Y KITTLER, J.V. (1982), *Pattern Recognition. A Statistical Approach*. Prentice Hall- Englewood Cliffs.

HILERA, J.R. Y MARTÍNEZ. V.J. (1995), *Redes Neuronales Artificiales. Fundamentos,*

modelos y aplicaciones. Madrid: RA-MA.

KOHAVI, R.; PROVOST, F. Y FAWCETT, T. (1998), The case against accuracy estimation for comparing induction algorithms. In C. S. Mellish, (ed), Proceedings of IJCAI-95, 1137-1143. Morgan Kaufmann.

LIM, T-S., LOH, W-Y. Y SHIH, Y-S. (1998), *An empirical comparison of decision trees and other classification methods.* Technical Report 979, University of Wisconsin, Madison.

MERZ, C.J., Y MURPHY, P.M. (1996), *UCI Repository of Machine Learning Databases.* Department of Information and computer science, University of Califormia, Irvine, CA. (http://www.ics.uci.edu/~mlearn/MLRepository.html)

MICHIE, D., SPIEGELHALTER, D.J. Y TAYLOR, C.C. EDITORES (1994), *Machine Learning, Neural and Statistical Classification*. Ellis Horwood.

MINSKY, M.L. Y PAPERT, S.A. (1969), *Perceptrons*. MIT Press, Cambridge, MA.

QUINLAN, J.R. (1986), Introduction of decision trees. Machine Learning, 1, 81-106.

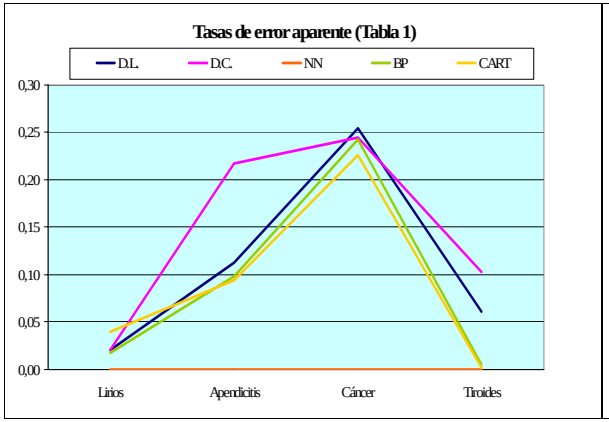
QUINLAN, J.R. (1993), C4.5: Programs for Machine Learning. San Mateo CA: Morgan Kaufmann

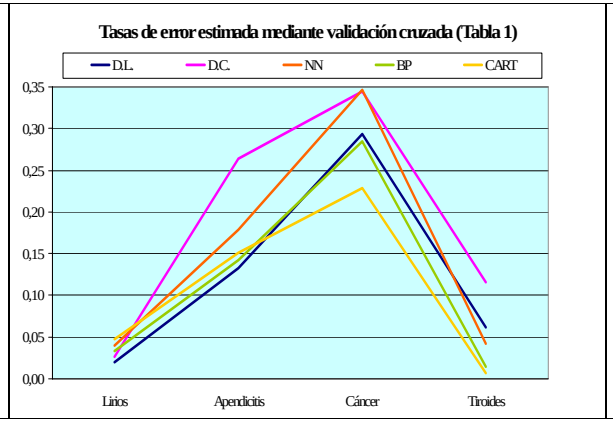
RIPLEY, B.D. (1996), *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

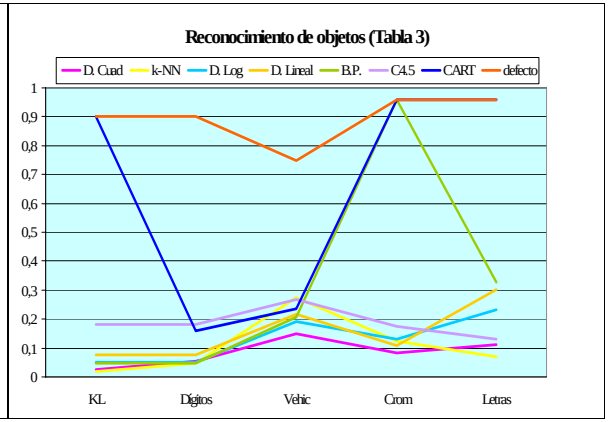
SALZBERG, S.L. (1999), *On comparing classifiers: a critique of current research and methods.* Data mining and knowledge discovery, 1, 1-12.

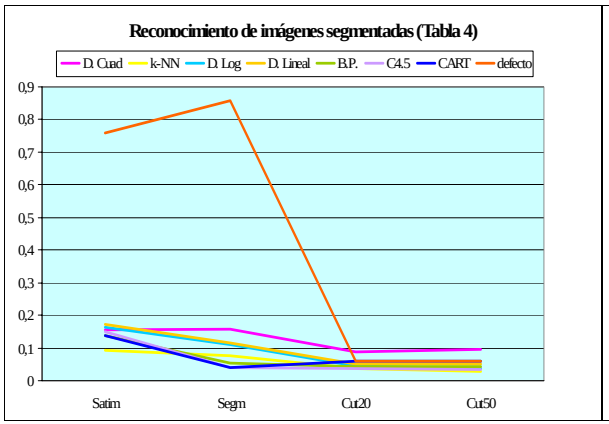
WEISS, S.M. Y KULIKOWSKY, C. (1991), *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann Publishers.

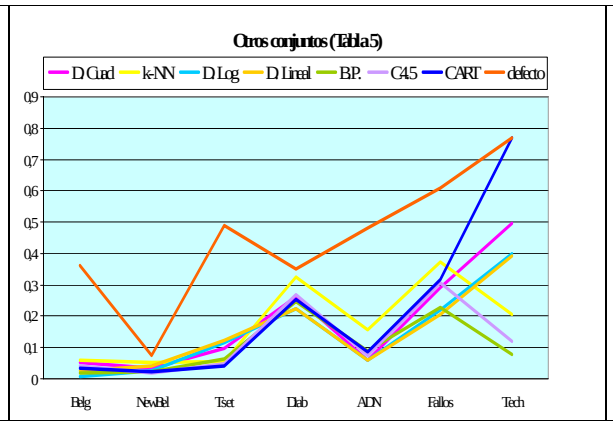
ANEXO.

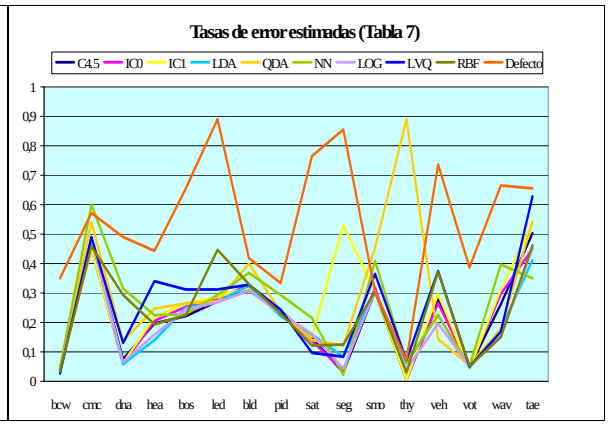

Tasas de error aparente (Tabla 1)
D.L.
D.C.
NN
BP
CART
0,30
0,25
0,20
0,15
0,10
0,05
0,00
Linios
Apendicitis
Cáncer
Tiroides


Tasas de error estimada mediante validación cruzada (Tabla 1)
D.L.
D.C.
NN
BP
CART
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
Linios
Apendicitis
Cáncer
Tiroides


Reconocimiento de objetos (Tabla 3)
D. Cuad
k-NN
D. Log
D. Lineal
B.P.
C4.5
CART
defecto
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
KL
Dígitos
Vehic
Crom
Letras


Reconocimiento de imágenes segmentadas (Tabla 4)
D. Cuad
k-NN
D. Log
D. Lineal
B.P.
C4.5
CART
defecto
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Satim
Segm
Cut20
Cut50


Otros conjuntos (Tabla 5)
D. Cuad
k-NN
D. Log
D. Lineal
B.P.
C4.5
CART
defecto
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Belg
NewBel
Tset
Dlab
ADN
Fallos
Tech


Tasas de error estimadas (Tabla 7)
C4.5
ICO
IC1
LDA
QDA
NN
LOG
LVQ
RBF
Defecto
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
bcw
cmc
dna
hea
bos
led
bld
pid
sat
seg
smo
thy
veh
vot
wav
tae