La predicción y la clasificación de datos en *marketing*.

Un análisis comparativo mediante técnicas multivariantes, árboles jerárquicos y redes neuronales

Jean-Pierre Lévy Mangin*, María Aranzazu Sulé Alonso**, Juan Salazar Clavel***

Recepción: marzo 7 de 2001
Aceptación: octubre 3 de 2001

* Universidad de Québec, Hull, Québec y La Cité, College of Applied Arts and Technology, Ottawa, Ontario, Canadá.
Correo electrónico:
jean-pierre_levy-mangin@uqah.uquebec.ca
**Universidad de León, España.
Correo electrónico: asule@gugu.usal.es
***Universidad de Québec, Hull, Québec, Canadá.

Resumen. Las técnicas mas utilizadas hasta el momento para la investigación y el análisis de datos en *marketing* eran casi exclusivamente multivariantes. Los desarrollos actuales de algunas de ellas, como los árboles jerárquicos, dan al *marketing* la precisión que le faltaba y permiten abordar la predicción y la clasificación con más seguridad. Las Redes Neuronales cubren los mismos tópicos con una precisión variable que se puede superar seleccionando mejor los datos y escogiendo los algoritmos adecuados.

Palabras clave: datos multivariantes, árboles jerárquicos, predicción, clasificación, redes neuronales, algoritmos.

The prediction and Classification of Data in *Marketing*: a Comparative Analysis of Multivariant Techniques, Hierarchical Trees and Neuronal Networks

Abstract. Until recently, the most widely used techniques for research and data analysis in marketing were almost exclusively multivariant. New developments with hierarchical trees give *Marketing* analysis the accuracy previously lacking and permit greater security with predictions and classification. Neuronal networks cover the same topics, with a variable accuracy that could be improved by selecting data better and choosing suitable algorithms. This paper will present a comparative study using three different techniques.

Key words: multivariate data, hierarchical trees, prediction, clasification, neural networks, algorithms.

Introducción

El *marketing* se distingue por ser una ciencia social que necesita utilizar técnicas que den los resultados más exactos posibles. Así como la segmentación y el posicionamiento en *marketing* quedan bastante bien cubiertos por técnicas como la clasificación y la segmentación jerárquica, el posicionamiento lo es por técnicas como el análisis por componentes principales, el análisis de correspondencias sencillas o múltiples y el escalamiento multidimensional. La predicción y la clasificación de datos evolucionan constantemente con técnicas a la vez tradicionales y otras de índole más moderna.

El propósito de este artículo es presentar comparativamente dichas técnicas y los resultados correspondientes.

El artículo pretende, con un mínimo de hincapié en la teoría, presentar los resultados comparativos llevados a cabo con varias técnicas de clasificación y predicción seleccionadas, entre ellas el análisis discriminante, la clasificación y regresión jerárquica, la segmentación jerárquica, y el análisis con redes neuronales.

Con ello se pretende presentar unos resultados y poder juzgar parcialmente –con el límite que puedan presentar los datos– por lo menos del fundamento legal de ciertas técnicas.

La base de datos trata de automóviles y ha sido tomada de la revista *Motor Trend* (15 de marzo de 1984) y queda

Figura 1. Matriz de clasificación y de predicción según un análisis discriminante.

Tabla de conteo de clasificación para CYLINDER

Actual	Predicción Ausente	cinco	cuatro	ocho	seis	tres	Total
Ausente	1	0	0	0	0	0	1
cinco	0	2	1	0	0	0	3
cuatro	0	16	177	1	4	6	204
ocho	0	0	0	96	5	0	101
seis	0	3	10	0	67	1	81
tres	0	0	0	0	0	4	4
Total	1	21	188	97	76	11	394

Reduccion en error de clasificación hacia X's = 85.7%

disponible también en las bases de datos del programa SPSS; contiene 407 casos de modelos de automóviles y presenta nueve variables de análisis de las que se utilizarán seis (millas por galón, capacidad de desplazamiento del motor en pulgadas cúbicas o motor, caballos vapor, peso en libras, aceleración y, como variable dependiente, la cilindrada con cinco categorías: 3, 4, 5, 6 y 8 cilindros).

El propósito de la investigación es identificar el método o la técnica que mejor prediga las categorías de pertenencia y que a la vez clasifique mejor los modelos, todo ello con el nivel de error más bajo.

Cabe hacer algunas advertencias: ciertas técnicas funcionan mejor con una base de datos amplia y con muchas variables; se trata de las técnicas asociadas a las redes neuronales que tienen que reconstituir el algoritmo o la fórmula a partir de los datos observables.

Puede existir un sesgo en el caso de valores ausentes, al tener en cuenta la manera en que sean considerados por el algoritmo o el sistema de análisis utilizado. El programa de redes neuronales utilizado los fija a cero y los demás programas suelen suprimir[1] el dato entero.

I. Análisis

La base de datos servirá para la clasificación y la predicción en relación con la variable dependiente *cilindrada*. Se utilizarán tres tipos de técnicas: el primero se fundamentará en técnicas multivariantes como el análisis discriminante; el segundo en técnicas jerárquicas de clasificación como la regresión, la clasificación jerárquica y la segmentación jerárquica (CHAID); el último tipo son técnicas heurísticas como las redes neuronales con funciones supervisadas de multicapas (Multi-Layer Perceptron)[2] y la de la Función de Base Radial.

1. Lo que ocurre con los datos de 11 a 15 y 18 para la variable millas por galón de la base de datos.
2. Mediante el uso del algoritmo de aprendizaje del Gradiante Descendente Conjugado (Descent Conjugate Gradiant).

Todas estas técnicas serán comparadas con el fin de prever y clasificar los modelos de automóviles en sus categorías respectivas. Las cinco variables independientes permitirán clasificar correctamente los casos en relación con las cinco categorías (3, 4, 5, 6 y 8 cilindros) de la variable dependiente cilindrada.

1. El análisis discriminante

El análisis discriminante constituye una técnica clásica y muy conocida del investigador, por ello no se la desarrollará teóricamente. El análisis realizado equivaldría, entonces, a una regresión múltiple cuya variable dependiente sería la variable cilindrada dividida en cinco categorías (3, 4, 5, 6 y 8 cilindros) y cuyas variables independientes serían millas por galón, capacidad de desplazamiento del motor en pulgadas cúbicas (motor), caballos vapor, peso en libras y aceleración.

La regresión discriminante presentará una matriz de predicción y de clasificación final que se puede observar en la figura 1.

El análisis discriminante ha llegado a clasificar correctamente 85.7% de los datos, lo que arroja un error del 14.3% en la predicción y en la clasificación, hay que añadir que donde existen valores ausentes (*missing values*) el programa ha suprimido el dato, por lo que los resultados finales pueden cambiar. Existen en total 407 datos y el programa recopila 394, por lo que se han suprimido 13, a los que habría que añadir los datos mal clasificados. De hecho, sobre un total de 407 el error es de 0.1498 lo que daría una clasificación correcta del 85.012% de los datos, muy próxima a lo que presenta la figura 1.

2. Los métodos jerárquicos

Los métodos jerárquicos como técnicas de análisis multivariante deben enmarcarse entre los métodos de dependencia pues, al igual que sucede con el resto de las técnicas de este tipo, se establece una distinción entre variables cuyo comportamiento se desea explicar, o variables dependientes, y aquellas que se utilizan para explicar las anteriores, denominadas explicativas o "predictoras", que son de naturaleza independiente.

La segmentación jerárquica pertenece a los métodos que utilizan un proceso o algoritmo basado en criterios que identifican grupos homogéneos de una población. En estos modelos los segmentos pueden formarse al considerar como criterio una única variable dependiente, en cuyo caso se denomina "monotética", o utilizar como criterio varias varia-

bles dependientes o "politéticas", como sucede en la segmentación canónica.

Por el contrario, las técnicas de clasificación o tipología, pertenecen a los modelos que no se basan en criterios y aunque se utilicen también para obtener subgrupos homogéneos, éstos se forman a partir de todas las variables recogidas en el análisis consideradas de forma conjunta. Por tanto, la tipología se trata de un método de agrupación ascendente que constituye una técnica descriptiva y composicional.

Figura 2. Matriz de predicción y de clasificación con variable dependiente ordinal.

Categoría Predecida	Categoría Actual					
	4	6	3	5	8	Total
4 cilindros	297	6	4	2	0	219
6 cilindros	0	77	0	1	0	78
3 cilindros	0	0	0	0	0	0
5 cilindros	0	0	0	0	0	0
8 cilindros	0	1	0	0	107	108
total	207	84	4	3	107	405

Resustitución

Riesgo estimado	0.0345679
Error estándar de riesgo estimado	0.00907758

La segmentación jerárquica es una técnica explicativa y descomposicional que utiliza un proceso de división secuencial, iterativo y descendente que, mediante la definición de una variable dependiente discreta o "variable a explicar", forma grupos homogéneos definidos específicamente mediante combinaciones de variables independientes explicativas en los que se incluyen la totalidad de los casos recogidos en la muestra.

La segmentación jerárquica también se conoce como "técnicas de árboles", porque utiliza una estructura de representación de la información similar a un gráfico de flujo o diagrama de árbol que ofrece una instantánea visual de los segmentos y patrones de los datos. La raíz del árbol es el conjunto de datos íntegro, los subconjuntos conforman las ramas del árbol. Un conjunto en el que se hace una partición se denomina nodo.

El árbol de decisión se construye al dividir el conjunto de datos en dos o más subconjuntos de observaciones a partir de los valores que toman las variables predictoras. Cada uno de estos subconjuntos vuelve después a ser particionado mediante el mismo algoritmo. Este proceso continúa hasta que no se encuentran diferencias significativas en la influencia de las variables de predicción de uno de estos grupos hacia el valor de la variable de respuesta. El proceso es secuencial e interactivo, pues recoge no sólo el efecto principal de las variables explicativas sobre la variable a explicar, sino también las interacciones entre las variables, es decir, la influencia que cada variable independiente produce en la variable dependiente en función de los valores que adoptan el resto de variables independientes contempladas en el análisis.

Los métodos jerárquicos se fundamentan principalmente en varias técnicas, la regresión y la clasificación jerárquica (C&RT) al usar el algoritmo de Breiman, y la segmentación jerárquica basada fundamentalmente en la clasificación automática según CHAID (Chi-square Automatic Interactive Detector), la clasificación exhaustiva también según CHAID y por fin la clasificación jerárquica según QUEST.

a) La regresión y la clasificación jerárquica

La regresión y clasificación jerárquica (ahora RCJ) constituye un método parecido al algoritmo AID que dicotomiza la muestra según el método de Breiman *et al.* (1984) y genera árboles de decisiones binarias. El propósito es realizar una partición de los datos en subconjuntos homogéneos con respecto a la variable dependiente. A cada partición de los datos se seleccionará la mejor variable independiente predictora (explicativa) que presentará la mejor dicotomización (para variables independientes continuas) o el mejor ajuste de categorías (para variables nominales u ordinales) con base en la mayor reducción de impureza, y se crearán dos nodos con el mejor predictor. El proceso se repetirá hasta que se alcancen los parámetros de paro. Contrariamente, con el algoritmo AID (Automatic Interative Detector), la dicotomización se efectúa de forma a maximizar la varianza entre la suma de los cuadrados del mismo grupo (ver Miquel *et al.*, 1997).

La variable dependiente en RCJ puede ser nominal, ordinal o continua. En nuestro ejemplo se ha considerado primero como ordinal y luego nominal.

El RCJ pasa por uno de los métodos más precisos de clasificación. La figura 2 muestra los datos mal clasificados y el error es de apenas 3.45%, es decir que los datos han sido clasificados correctamente a 96.55%, un índice muy alto.

La figura 3 muestra el árbol de clasificación con los distintos predictores, nodos y grupos.[3]

En este caso se ha considerado la variable dependiente *cilindros* como ordinal, es decir, que las categorías son continuas

3. La variable que mejor clasifica el número de cilindros según la RCJ es la variable explicativa motor (capacidad de desplazamiento del motor en pulgadas cúbicas) en el punto 159,500 pulgadas con los valores superiores en la rama derecha e inferiores en la rama izquierda. El algoritmo sigue clasificando con la misma variable explicativa para valores superiores e inferiores a 259,000 pulgadas cúbicas.

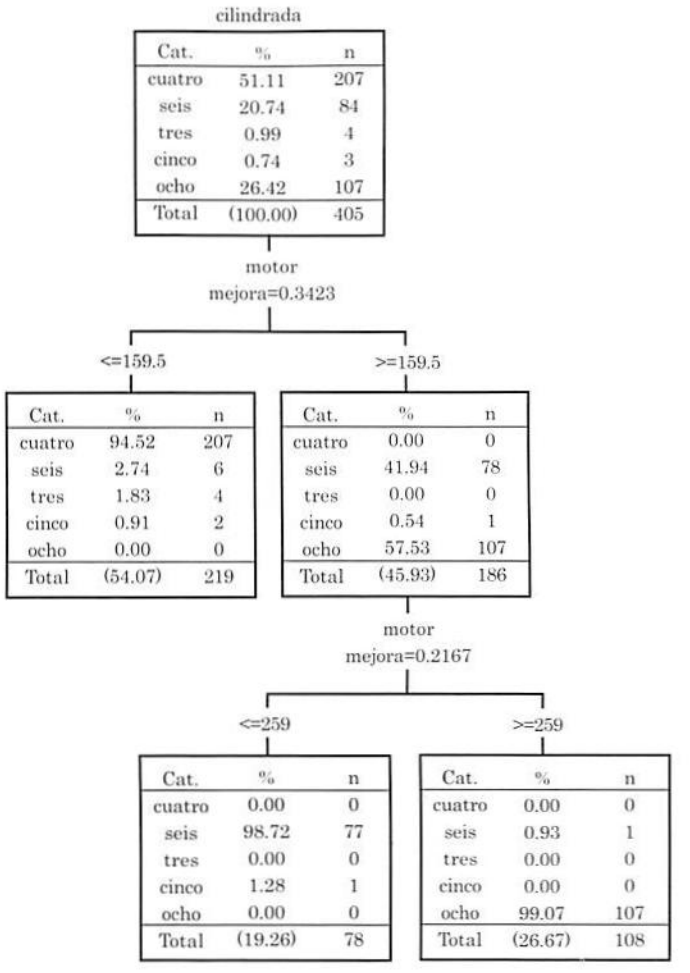
y que 6 cilindros es inferior a 8 y 8 superior a 6, etcétera.

También se puede considerar la variable dependiente cilindros como nominal y en ese caso se desubica la relación ordinal entre las categorías. La matriz de predicción y clasificación arrojará los resultados de la figura 4.

Se puede apreciar el ajuste, ligeramente distinto, se ha añadido un dato ausente y se ha clasificado mal un modelo de ocho cilindros en cuatro. El error en este caso es de 4.6%, por lo que los datos son clasificados correctamente al 95.33%.

El árbol de decisión será el mismo que el de la figura 3 con la variante de los dos casos mencionados.

Figura 3. Árbol de decisión de RCJ según el algoritmo C&RT con variable dependiente ordinal.*



* La variable que mejor clasifica el número de cilindros según la RCJ es la variable explicativa *motor* (capacidad de desplazamiento del motor en pulgadas cúbicas) en el punto 159,500 pulgadas con los valores superiores en la rama derecha e inferiores en la rama izquierda. El algoritmo sigue clasificando con la misma variable explicativa para valores superiores e inferiores a 259,000 pulgadas cúbicas.

Figura 4. Matriz de predicción y de clasificación con variable dependiente nominal.

Categoría Predecida	Categoría Actual						
	ocho	cuatro	seis	ausente	tres	cinco	Total
ocho	106	1	1	0	0	0	108
cuatro	0	207	6	1	4	2	220
seis	0	3	74	0	0	1	78
ausente	0	0	0	0	0	0	0
tres	0	0	0	0	0	0	0
cinco	0	0	0	0	0	0	0
total	106	211	81	1	4	3	406

Resustitución

Riesgo estimado 0.046798

Error estándar de riesgo estimado 0.010482

b) La segmentación jerárquica

Uno de los primeros algoritmos empleados en segmentación jerárquica fue el AID o Detección Automática de Interacciones creado por Sonquist y Morgan de la Universidad de Michigan a principios de los años sesenta, que se utilizó mucho en la década de los setenta y principios de los ochenta hasta que surgió el CHAID.

La segmentación jerárquica (ahora SJ) y más particularmente el algoritmo CHAID se debe a Kass (1980) y su desarrollo posterior a Jay Magidson (1993). Contrariamente al algoritmo AID, el CHAID utiliza variables nominales, ordinales y continuas, y actúa por disgregación politómica (y no dicotómica) de las categorías o clases.

De manera general, la SJ permite descomponer una base de datos en varios grupos con base en el mejor predictor de la variable dependiente (en función de la prueba de la χ^2 más significativa).

La SJ permite determinar la mejor partición de cada nodo y aparejar clases o categorías de las variables independientes predictoras. Se repite el proceso hasta que no queden más parejas significativas.

Seguidamente se vuelve a seleccionar el nuevo mejor predictor y el grupo (nodo) será dividido en dos o más clases. Se repetirá el proceso hasta que se llegue a la regla de fin.

Dicho rápidamente, se podría comparar la SJ a una regresión de tablas cruzadas desde la más significativa a la menos. El análisis de correspondencias constituirá el complemento natural de esta técnica (ver Salazar *et al.*, 1997).

La crítica que se hizo a Kass es que nada garantiza la partición óptima de cada grupo (en cada nodo), para ello habría que realizar una investigación exhaustiva de todos los grupos posibles hasta que quede el grupo más significativo, cosa que no realiza CHAID pero sí el algoritmo propuesto por Biggs *et al.* en 1991 con el nombre de Exhaustive CHAID. El algoritmo actúa de la misma manera que el anterior, pero también a la vez de manera recursiva.

Figura 5. Matriz de predicción y de clasificación de SJ según CHAID.

Categoría Predecida	Categoría Actual						
	ocho	cuatro	seis	ausente	tres	cinco	Total
ocho	106	1	6	0	0	0	113
cuatro	0	189	0	1	4	2	196
seis	0	21	75	0	0	1	97
ausente	0	0	0	0	0	0	0
tres	0	0	0	0	0	0	0
cinco	0	0	0	0	0	0	0
total	106	211	81	1	4	3	406

Resustitución

Riesgo estimado	0.08867
Error estándar de riesgo estimado	0.0141079

La SJ también constituye una técnica de clasificación y se podrá observar en la figura 5 cómo son clasificados los modelos de automóviles.

Se observa que la clasificación según SJ arroja un error del 0.08867, es decir, que los datos son clasificados correctamente al 91.133%, que es inferior a la RCJ, aunque muy aceptable.

Se puede observar también en la figura 6 que el árbol politomiza la muestra y no la dicotomiza.

También se llevó a cabo una investigación con CHAID exhaustivo que no dio resultados satisfactorios. Con la variable cilindrada considerada como ordinal arrojó un error de clasificación de 0.1777 y de 0.1724 con la misma variable considerada como nominal. Los árboles eran parecidos al de CHAID con la diferencia que había un nivel más (dos en total).

c) La clasificación y la regresión jerárquica según QUEST

El algoritmo de RCJ propuesto por Breiman *et al.* (*op. cit.*) suele seleccionar más variables discretas que permiten llevar a cabo más particiones del árbol jerárquico, además el algoritmo de RCJ es gran consumidor de tiempo CPU de ordenador.

El método QUEST para Quick, Unbiai-sed, Efficient Statistical Tree descrito por Loh y Shih (1997) constituye un método de partición binaria de una base de datos y, contrariamente a la RCJ y a la SJ según CHAID, selecciona primero la variable predictora y luego efectúa la partición (la RCJ y la SJ lo llevan a cabo de manera simultánea).

Para cada partición, la asociación entre la variable predictora y la variable dependiente es llevada a cabo por

Figura 6. Árbol de SJ según CHAID.*

cilindrada

Cat.	%	n
tres	0.99	4
ocho	26.11	106
cinco	0.74	3
cuatro	51.97	211
seis	19.95	81
ausente	0.25	1
Total	(100.00)	406

motor

valor-P=0.0000, Chi-cuadrado=644.7775, df=15

(4,140)

Cat.	%	n
tres	2.04	4
ocho	0.00	0
cinco	1.02	2
cuatro	96.43	189
seis	0.00	0
ausente	0.51	1
Total	(48.28)	196

(140.250)

Cat.	%	n
tres	0.00	0
ocho	0.00	0
cinco	1.03	1
cuatro	21.65	21
seis	77.32	75
ausente	0.00	0
Total	(23.80)	97

(250.318)

Cat.	%	n
tres	0.00	0
ocho	86.79	46
cinco	0.00	0
cuatro	1.89	1
seis	11.32	6
ausente	0.00	0
Total	(13.05)	53

(318.455)

Cat.	%	n
tres	0.00	0
ocho	100.00	60
cinco	0.00	0
cuatro	0.00	0
seis	0.00	0
ausente	0.00	0
Total	(14.78)	60

* Con CHAID, la variable que mejor clasifica a la *cilindrada* es la variable explicativa *capacidad de desplazamiento del motor en pulgadas cúbicas* (motor) que se divide en cuatro ramas, la primera que va hasta 4,140 pulgadas cúbicas, la segunda hasta 140,250, etcétera.

la prueba F de Anova, la prueba de Levene para las variables ordinales y continuas y la prueba de la χ^2 de Pearson para las variables nominales.

Se aplica el análisis discriminante cuadrático para identificar el mejor punto de partición de la variable predictora, proceso que se repetirá hasta llegar a la regla de fin (inexistencia de predictores significativos).

Si la variable dependiente es única, siempre deberá ser nominal, con ello este algoritmo es superior y debería dar mejores resultados que la RCJ.

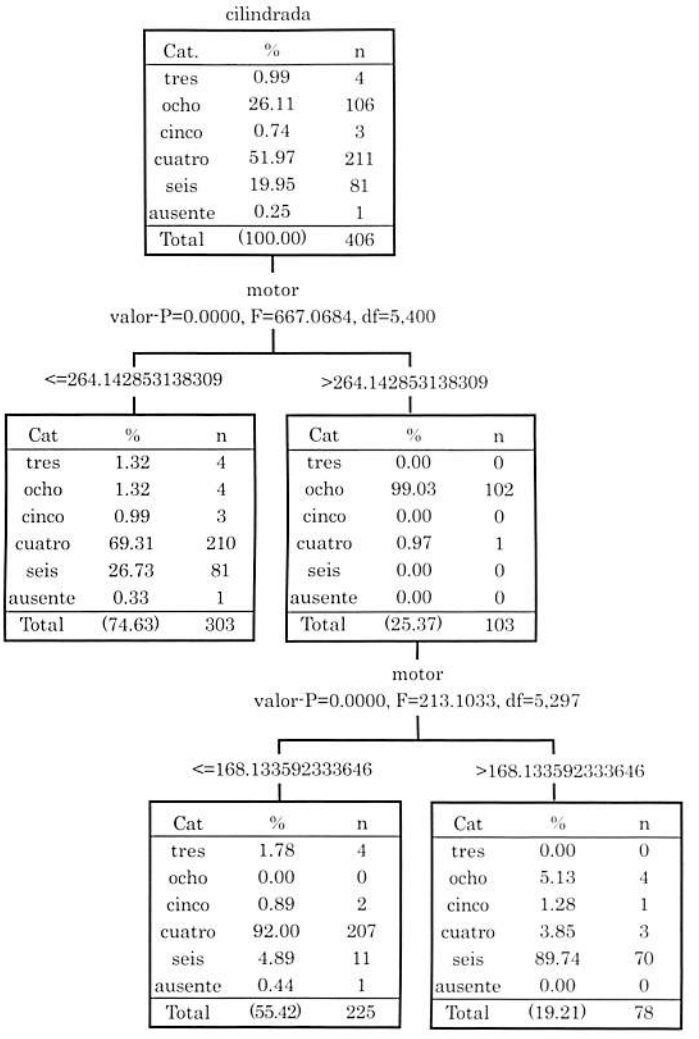
Se puede observar que QUEST, aunque se presente como un programa superior, no mejora la clasificación anterior de la RCJ. El error de clasificación, como se observará en la figura 8a, es de 0.0665, lo que significa que los datos son clasificados correctamente al 93.35%, resultado superior a la SJ con CHAID pero inferior a la RCJ en general.

Existen otros métodos de clasificación, como por ejemplo los algoritmos FACT (Loh y Vanichestakul, 1988) o THAID (Morgan y Messenger, 1973) y con CHAID se puede decir que son métodos de descomposición multiniveles y no meramente métodos de descomposición binarios. Una controversia ha sido declarada por Loh y Shih en 1997, según la cual hay pocas ventajas en llevar a cabo descomposiciones multiniveles, es decir, politomizando la muestra, en vez de descomposiciones binarias que según estos autores clasifican mejor. También según ellos cualquier descomposición politómica puede ser llevada a cabo por una serie de descomposiciones binarias o dicotómicas profundizando el árbol.

Habría que ponderar un poco estas consideraciones, si de clasificación se trata o de encontrar el mejor predictor y el mejor punto de partición se puede escoger cualquier método de RCJ. Si se trata de identificación de segmentos de mercado o de jerarquizar la importancia de ciertas tablas cruzadas con respecto a otras, la SJ cobra toda su importancia como técnica explicativa y descriptiva (con el análisis de correspondencias).

Como bien decía Richard Bagozzi (1993), "por fin se va a poder parar de estructurar a mano infinidad de tablas

Figura 7. Árbol de RCJ según QUEST.*



* El árbol jerárquico según QUEST presenta la variable *cilindrada* como variable dependiente, la variable explicativa que mejor clasifica es capacidad de desplazamiento del motor en pulgadas cúbicas (*motor*) en el punto 264.143 pulgadas con los valores superiores en la rama derecha e inferiores en la rama izquierda. El algoritmo sigue la clasificación con la misma variable explicativa para valores superiores e inferiores a 168.134 pulgadas cúbicas.

Figura 8a. Matriz de predicción y de clasificación según QUEST.

Categoría Predecida	Categoría Actual						
	ocho	cuatro	seis	ausente	tres	cinco	total
ocho	102	1	0	0	0	0	103
cuatro	0	207	11	1	4	2	225
seis	4	3	70	0	0	1	78
ausente	0	0	0	0	0	0	0
tres	0	0	0	0	0	0	0
cinco	0	0	0	0	0	0	0
total	106	211	81	1	4	3	406

Resustitución

Riesgo estimado	0.0665025
Error estándar de riesgo estimado	0.0123655

cruzadas según el nivel de significación, ya existe un programa que las estructura y las jerarquiza automáticamente".

3. Las redes neuronales y los elementos fundamentales de la estructura de una red con aprendizaje supervisado

Las redes neuronales son sistemas de procesamiento inspirados en el funcionamiento del cerebro humano, compuestos por un conjunto de neuronas que procesan e intercambian información. El aprendizaje se basa fundamentalmente en la experiencia.

Las neuronas de una red se estructuran en capas de forma que la neurona de cada capa está interconectada con las neuronas de la capa siguiente.

La neurona podría ser definida como un procesador elemental caracterizado por un estado interno, por señales de entrada y por una función de transferencia.

Las neuronas están formadas en capas sucesivas y cada capa está compuesta por un conjunto de neuronas sin conexión. Cada neurona recibe señales a través de las conexiones con las neuronas de las capas anteriores.

Cada vez que una neurona manda información a otra de la capa superior, una función opera los valores recibidos con los pesos asignados (llamados pesos sinápticos) en cada una de las conexiones y si el resultado de la operación supera un umbral, la neurona se activa y emite una señal a las neuronas de la capa siguiente.

Los pesos sinápticos almacenan la memoria a largo plazo y determinan el carácter activador o inhibidor de las entradas de datos.

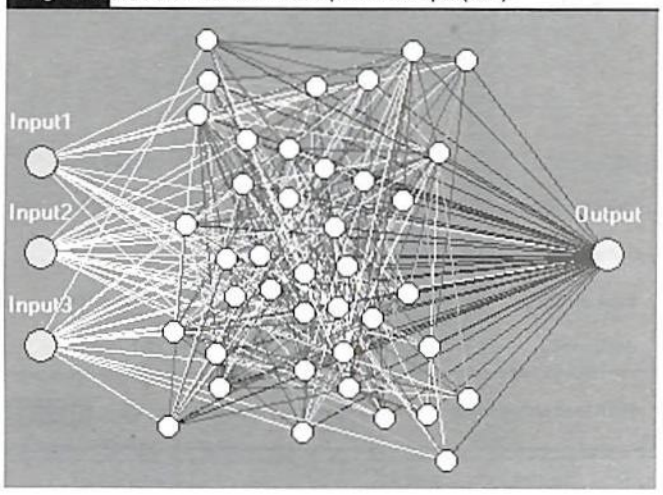
Sobre los datos de entrada actúa una matriz de pesos iniciales (comúnmente llamada W) obteniéndose un vector de valores (V) que se filtra a través de una función que puede ser lineal, sigmoidal o tangente, los datos resultantes corresponden a la capa intermedia u oculta.

El proceso se repite desde la capa intermedia a la salida de datos mediante una nueva matriz de pesos (W´) hasta dar con el patrón de salida deseado.

El aprendizaje supervisado trata de encontrar en los datos de salida el patrón deseado a priori.

El problema en redes neuronales se plantea en términos de buscar el vector de pesos W que minimice una determinada función de error (diferencia entre la salida y el patrón)

Figura 8b. Redes Neuronales con Perceptrón Multicapas (MLP).



teniendo en cuenta la información proporcionada por la muestra de aprendizaje. Se trata, pues, de un problema clásico de optimización y para ello se usarán distintos algoritmos de optimización más o menos complejos y que por supuesto son múltiples (al igual que las estructuras de las neuronas son también múltiples).

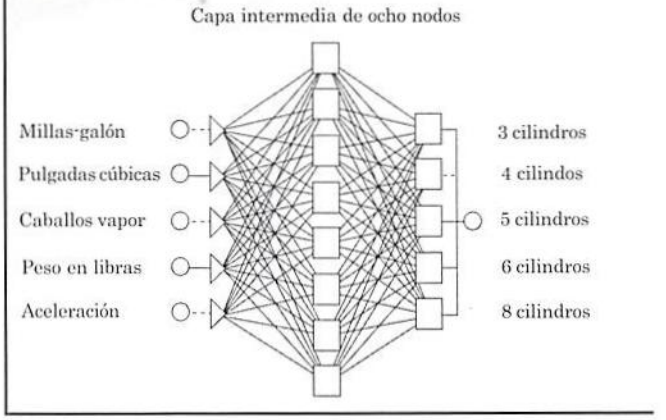
a) El análisis del Perceptrón Multicapas (MLP)

El modelo de ARN (Análisis de Redes Neuronales) más conocido como el Perceptrón Multicapas se fundamenta en varios paradigmas, el principal y más utilizado es el del aprendizaje por retropropagación de error (Back-Propagation) que constituye una presentación general de la arquitectura de estos modelos (Patterson, 1996; Haykin, 1994 y Fausett, 1994). Se presentará también el algoritmo por gradiente descendente.

–El algoritmo por retropropagación

El Perceptrón Multicapas por retropropagación del error es posiblemente la arquitectura más popular y fue presentada en sus principios por Rumelhart y McClelland (1986). Las unidades de neuronas se presentan en varias capas intermedias en un modelo con entradas (*inputs*) y salidas (*outputs*). Las unidades de neuronas intermedias harán un cálculo ponderado de la suma de las entradas, lo que dará lugar a la presentación de una matriz de pesos y se obtendrá un vector de valores que será filtrado de nuevo por una función (sigmoíde, arcotangente u otra) para dar como resultado la salida. Si esta salida no correspondiera al patrón de salida y que el error fuera muy importante se volverían a introducir los datos en la red, por eso este algoritmo se denomina de retropropagación y el proceso es de aprendizaje con los datos de la muestra. El aprendizaje supervisado en este caso es por corrección del error, cuya función sería la del error cuadrático medio.

Figura 9. Arquitectura MLP con capa intermedia de ocho nodos



La salida (*output*) de cualquier nodo de la capa intermedia es estimada por la siguiente función $Z_j = \varphi\Sigma_{i=0}^{n-1} W_{ij}\, X_i$ donde φ corresponde a la función de transferencia del nodo, W_{ij} corresponde a los pesos entre los nodos *i* y *j*; Z_j es la salida del nodo *j*.

La salida es comparada al patrón, la diferencia es el error. El proceso se repite capa por capa hasta que la salida real sea lo más próxima posible a la salida deseada y que el error sea mínimo. Sin embargo, existen otros algoritmos más modernos y más rápidos.

–El algoritmo del Gradiente Descendente Conjugado (*Conjugate Gradient Descent*).

Este algoritmo funciona de manera diferente al de retropropagación, es más moderno y llega al mínimo muy rápidamente a partir de una configuración de partida. El problema es que se puede tratar de un mínimo local que podría pasar por un mínimo óptimo. Para ello hay que repetir varias veces la misma operación y comparar los errores de cada ejecución con el error más bajo.

Estos algoritmos buscan una dirección "sensible" hacia la cual dirigirse en un plano multidimensional, identifican el mínimo, proyectan una línea y se dirigen a él (Bishop, 1995; Sheperd, 1997). Para no llegar a mínimos locales cuando se selecciona una dirección hay que seleccionar, como decía Bishop, direcciones "comúnmente orientadas" (*conjugate*) pero sin influencia entre ellas.

La primera etapa de la utilización del ARN será constituida por el diseño de la red; en cualquier arquitectura la primera capa de entrada será constituida por las cinco variables independientes (millas por galón, capacidad de desplazamiento del motor en pulgadas cúbicas, caballos vapor, peso en libras, aceleración, cilindrada), la segunda (de análisis) será diseñada por el investigador o por el algoritmo de manera automática y la tercera capa será constituida por los cinco valores del patrón o de la variable dependiente, es decir, el número de cilindros: 3, 4, 5, 6 u 8.

La o las capas (1 o 2) intermediarias de neuronas albergan funciones de transformación de los datos de entrada que intentarán hacer coincidir los resultados con el patrón de salida.

La segunda etapa será constituida por el aprendizaje de la red y consiste en utilizar un algoritmo que ajustará los pesos (sinápticos) de las conexiones entre las neuronas. Los casos de aprendizaje (*training*) permitirán de manera recurrente conseguir el objetivo, disminuir la función de error y conseguir la mayor eficacia de configuración de los pesos sinápticos de la red.

Los casos de validación vendrán a confirmar el aprendizaje realizado por la red, lo que permitirá clasificar una base de datos en función del patrón de salida deseado, que será dividida en tres: los casos de aprendizaje, los de validación y los de análisis.

La muestra se compone de 407 casos divididos en 204 de aprendizaje, 102 de validación y 101 de estimación. La partición ha sido realizada arbitrariamente por el ordenador y se estudiarán los casos de aprendizaje, de validación y de estimación.

b) El ARN multicapas o MLP

El modelo de ARN de multicapas utilizado se fundamenta en el algoritmo Perceptrón Multicapas con paradigma "Back-Propagation", se trata de la forma más común de modelos neuronales y una presentación general de la arquitectura de estos modelos.

Para solucionar este problema de clasificación se ha llevado a cabo un ARN multicapas utilizando una capa intermedia de ocho nodos (generada por el algoritmo, se realizó otro intento manual con dos capas de neuronas intermedias con menos precisión), la función de transformación de los nodos de la capa intermedia es una función tangente y el algoritmo de aprendizaje es el del Gradiante Descendente Conjugado (Descent Conjugate Gradiant). Para ello se utilizó el programa *Neural Connection de SPSS*. [4]

En la figura 9 se puede apreciar la arquitectura multicapas utilizada. La capa de neuronas de entrada se compone de las cinco variables independientes, la capa intermedia (de análisis) de ocho neuronas y la capa de salida incluye las cinco categorías de la variable dependiente *cilindrada*.

La figura 10 presenta la matriz de clasificación y de predicción que arroja un ajuste de datos del 96.08%. Este ajuste no es el final, debido a que se trata solamente de los casos de aprendizaje.

4. El programa *Neural Connection* utiliza por defecto el algoritmo del Gradiante Descente Conjugado, es también el más preciso.

Figura 10. Matriz de clasificación y predicción para un ARN con una capa intermedia de ocho nodos (aprendizaje).

Verdadero	Predecido					
	ocho	cuatro	seis	ausente	tres	cinco
ocho	75	0	1	0	0	0
cuatro	1	80	1	0	0	0
seis	0	2	41	0	0	0
ausente	0	1	0	0	0	0
tres	0	2	0	0	0	0
cinco	0	0	0	0	0	0

Número total de ejemplos: 204
Número total de objetivos: 204
Total correcto: 196
Porcentaje correcto: 96.08%

La tabla 1 muestra que los casos de aprendizaje quedan muy bien ajustados (96.08 %) aun cuando los casos de validación (que permiten la generalización) y de estimación lo sean menos, lo que significa que podría existir sobre-ajuste (*overfitting*) de los datos de aprendizaje. Ello se puede corregir y mejorar al escoger en parte los datos, es decir, transferir casos de aprendizaje a la validación y a la estimación y obtener aproximadamente el mismo porcentaje de casos clasificados correctamente. Si se asume el total, se considerará que el porcentaje de casos clasificados correctamente es del 91.15% para este tipo de ARN (multicapas con paradigma "Back-Propagation").

c) El ARN según modelos de Función Base Radial (RBF)

Así como en el Perceptrón Multicapas todas las capas tienen la misma estructura (lineal), en la Función de Base Radial (RBF) la capa intermedia tiene precisamente una estructura radial.

Los modelos de la RBF tienen sus orígenes en los escritos de Powell (1985), Broomhead y Lowe (1988), Moody y Darkin (1989) y Haykin (1994) que dividen los datos en segmentos correspondientes a regiones del espacio decisional definidas por las características de los datos y determinadas por funciones de base radial delimitadas por círculos al interior del espacio decisional.

La RBF es una función supervisada con patrón de salida que realiza clasificaciones (o previsiones) a partir de elipses e hiperelipses que parten el espacio de entrada de datos. Las elipses son definidas por la función radial siguiente: $f_i = \sum_{j=1}^{m} \lambda_{ij} \phi(\|x - y_j\|)$, donde $\|...\|$ representa una medida de distancia entre las entradas o *inputs* X y un centro de gravedad Y posicionado en el espacio de entrada de datos. Las hiperelipses son definidas por las funciones radiales ϕ del tipo $\phi(\|x - y\|)$.

La función respuesta de una unidad radial sencilla es una función de Gauss en forma de campana cuya pendiente puede ser definida por sus pesos y sus umbrales.

Figura 11. Función de Base Radial (RBF) en forma de campana (función de Gauss).

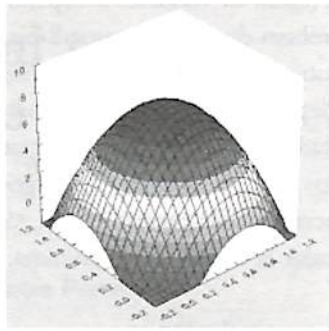


Figura 12. Matriz de clasificación y predicción para un ARN de Función Base Radial (casos o datos de aprendizaje).

Verdadero	Predecido				
	ocho	cuatro	seis	ausente	tres
ocho	76	0	0	0	0
cuatro	1	81	0	0	0
seis	0	1	42	0	0
ausente	0	0	0	1	0
tres	0	2	0	0	0
cinco	0	0	0	0	0
	0	0	0	0	0

Número total de ejemplos: 204
Número total de objetivos: 204
Total correcto: 200
Porcentaje correcto: 98.04%

Tabla 1. Casos clasificados erróneamente y porcentajes correctos de clasificación para un MLP con una capa oculta de ocho nodos intermedios.

Aprendizaje	Validación	Estimación	Total
8 casos	14 casos	14 casos	36 casos
96.08 %	86.27 %	86.14 %	91.15 %

Tabla 2. Casos clasificados erróneamente y porcentajes correctos de clasificación para un ARN de Función Base Radial.

Aprendizaje	Validación	Estimación	Total
4 casos	14 casos	15 casos	33 casos
98.04 %	86.27 %	85.15 %	91.89 %

La RBF presenta ciertas ventajas con respecto al ARN con multicapas y es precisamente que se puede modelizar con sólo una capa intermedia en vez de varias. Además, se puede decidir antes de la optimización lineal y con antelación el número de unidades radiales así como los centros y las desviaciones que serán aplicados a la capa de salida de una red de RBF.

En la fase de aprendizaje de los datos hay que fijar los centros de gravedad y las desviaciones de las unidades radiales con el propósito de optimizar la capa de salida lineal.

Existen dos técnicas de fijación de los centros de gravedad, por polimuestreo (Haykin, 1994) o a través del algoritmo K-medias (Bishop, 1995).

El algoritmo RBF presenta también otra ventaja, es sin duda mucho más rápido y nunca llega a una solución local como el ARN.

Contrariamente al modelo MLP, los pesos sinápticos de la capa de entrada (*inputs* de variables independientes) del modelo RBF no cambian durante el aprendizaje de los casos y la capa oculta utiliza un algoritmo K-medias para agrupar las variables independientes según sus similitudes.

En este apartado se pueden hacer exactamente los mismos comentarios que en el apartado anterior, el ARN por RBK puede a veces ser más preciso que el ARN por MLP, en este caso particular la diferencia es mínima.

Implicaciones y conclusiones

Se ha podido observar que desde los métodos estadísticos multivariantes ha habido cierta evolución en la precisión de la predicción y la clasificación de datos. Cualquier método con una precisión de clasificación superior al 85% debe ser considerado.

Los métodos multivariantes suelen ser invariables, es decir, que a partir de unos datos determinados la herramienta de cálculo es fija y presentará siempre los mismos resultados.

La herramienta puede, sin embargo, ser más moldeable con los métodos de árboles jerárquicos, pues se pueden sacar más clases al modificar ciertos puntos de partición. El método mostrará si la predicción es buena o no.

Estas técnicas han arrojado muy buenos resultados en esta investigación, en otras han sido quizá menos precisas, y aunque los valores por defecto suelen dar buenos resultados, a veces se pueden mejorar cambiando ciertos parámetros. Estas técnicas ofrecen bastantes métodos alternativos que se pueden probar en los datos y a veces alguno es más adaptado a cierto tipo de problemas. En general, las técnicas de árboles jerárquicos son bastante precisas y fiables.

Los ARN presentan las técnicas más moldeables y aunque los valores por defecto arrojen resultados aceptables, la selección propia y justificada de algoritmos y la manipulación de parámetros permiten a menudo ajustar los resultados mucho más eficazmente que al usar los valores por defecto (algoritmos y parámetros). A veces da la impresión de que la utilización de estas técnicas y la solución de ciertos problemas sea más bien una cuestión de destreza, el ajuste de ciertos datos se ha vuelto muy sensible. Es quizá la parte más interesante de los ARN, moldear la herramienta para llegar a resultados que puedan ser muy precisos, incluso superiores a las demás técnicas.

Cabe añadir que algunos programas hacen que sus algoritmos sean de uso mas amigable y permitan hallar más fácilmente una convergencia hacia un resultado significativo y satisfactorio.

Los programas abundan y la literatura se encarga de presentarlos y analizarlos, cada sistema estadístico de análisis de datos suele presentar sus propios programas por lo que el lector deberá hacerse una opinión por sí mismo. e/g/o

Bibliografía

Bagozzi, R. (1993). "The CHAID Approach to Segmentation Modeling", *Handbook o/Marketing Research*. Blackwell, Cambridge, Massachussets.

Biggs, D.; B. de Ville and E. Suen (1991). "A Method of Choosing Multiway Partitions for Classifications and Decision Trees", *Journal of Applied Statistics*, No. 18.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Breiman, L.; J.-H. Friedman; R. Olshen and C.J. Stone (1984). *Classification and Regression Trees*. Belmont , CA: Waldsworth.

Broomhead, D. S. and D. Lowe (1988). "Multivariate Functional Interpolation and Adaptative Networks", *Complex Systems*, pp. 321-355.

Fausett, L. (1994). *Fundamentals of Neural Networks*. Prentice Hall , New-York.

Haykin, S. (1994). *Neural Networks, a Comprehensive Foundation*. Macmillan, New York.

Kass, G. (1980). "An Exploratory Technique for Investigating Large Quantities of Categorical Data", *Applied Statistics*.

Loh, W. and Shih (1997). "Split Selection Methods for Classification Trees", *Statistica Sinica*.

Loh, W. and Vanichestakul (1988). "Fact and Some Classification Trees Programs", *Statistica Sinica*.

Magidson, J. (1993). "The *CHAID* Approach to Segmentation Modeling", *Handbook of Marketing Research*. Bagozzi, R.; Blackwell Cambridge, Massachussets.

Miquel, S.; E. Bigné; J. P. Lévy; A. C. Cuenca y M. J. Miquel (1997). *Investigación de mercados*. McGraw Hill.

Moody, J. and C. J. Darkin (1989). "Fast Learning in Networks of Locally Tuned Processing Units", *Neural Computation*. 1 (2): 281-294.

Morgan and Messenger (1973). "THAID, Sequential Analysis Program for the Analysis of Nominal Data". Document 5, *Robust Statistical Procedures*.

Patterson, D. (1996). *Artificial Neural Networks*. Prentice Hall.

Powell, T. (1995). *Number Crunching is Fun: Neural Computing in Market Research*. Marketing Computers.

Powell, M. J. D. (1985). *Radial Basis Functions for Multi-Variable Interpolation, a Review*. Department of Applied Mathematics and Theorical Physics, University of Cambridge.

Rumelhart, D. E. and J. L. McClelland (1986). "Parallel Distributing Processing". I-II MIT Press.

Salazar C., J.; A. Lambert y J. P. Lévy Mangin (1997). "La segmentación jerárquica y el posicionamiento mediante el uso conjunto del algoritmo Chaid y del análisis de correspondencias: una aportación metodológica, *ESIC-Market*, No. 97.

Sheperd, A. J. (1997). *Second Order Methods for Neural Networks*. Springer Verlag, New York.