Leandro Kovalevski
Paula Macat
Instituto de Investigaciones Técnicas y Aplicadas, Escuela de Estadística, Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario, Rosario, Argentina.

ALTERNATIVAS NO PARAMÉTRICAS DE CLASIFICACIÓN MULTIVARIADA

I. Introducción

El problema de la clasificación de individuos en poblaciones o grupos conocidos y de la caracterización de los mismos en base a un conjunto de variables medidas sobre los individuos, es de gran interés en estadística. Por esta razón se han desarrollado diversas técnicas para cumplir con este objetivo. Algunas de las más conocidas son:

- Análisis discriminante lineal
- Análisis discriminante cuadrático
- Regresión logística

El Análisis Discriminante es una de las técnicas más utilizadas para clasificación. De forma general, es un método que permite analizar las diferencias entre grupos de individuos previamente definidos, considerando alguna función lineal o cuadrática de las variables, y emplear luego esa función para predecir la pertenencia de una nueva observación a alguno de los grupos. El Análisis Discriminante resulta óptimo cuando las variables provienen de una distribución normal multivariada con igual variancia dentro de cada grupo (homocedasticidad), y sus resultados pueden no ser válidos ante la presencia de algunos pocos valores extremos (Khattree, R. Naik, D. 2000).

Dado que el requerimiento de normalidad y homocedasticidad no se cumple con frecuencia, es necesario utilizar técnicas que no requieran de tal supuesto, como la regresión logística (Barajas ,2007; Usuga, 2006; y Castrillón, 1998) o los métodos no paramétricos de clasificación, entre ellos: el método de los k vecinos más cercanos; el basado en núcleos Kernel; los árboles de decisión y las redes neuronales artificiales (López, M, et al., 2007; Härdle & Müller, 1997; Hechenbichler & Schliep, 2004; Cuadras, Carlos M., 2010; Manel Martínez Ramón, 2008).

En este trabajo se plantea como objetivo la aplicación de la técnica multivariada de clasificación no paramétrica: Árboles de Clasificación y Regresión, en adelante CART (de sus siglas en inglés, Classification And Regression Trees), perteneciente a la familia de las técnicas Árboles de decisión (Breiman et al. 1984), y su comparación con otros métodos como el de vecinos más cercanos en un contexto donde los métodos clásicos no son los más adecuados para el análisis debido a las características de las variables estudiadas.

II. Árboles de Clasificación y Regresión

II.1. Marco Teórico

Frecuentemente la investigación estadística se ve enfrentada a manipular grandes cantidades de datos complejos que incluyen un gran número de variables, de los cuales es necesario obtener información, encontrar patrones y definir tendencias. Con este propósito Sonquist, Baker y Morgan, (1971) propusieron el programa AID (Automatic Interaction Detection), el cual representa uno de los primeros métodos de ajuste de los datos basados en modelos de árboles de clasificación (Hadidi, 2003). En 1980, Kass propone un algoritmo recursivo de clasificación no binaria llamado CHAID (Chi Square Automatic Interaction Detection). Otros métodos más recientes son: FIRM (Formal Inference-based Recursive Modeling) propuesto por Hawkins (Hadidi, 2003); y MARS (Multivariate Adaptive Regression Splines), propuesto por Friedman en el año 1991. Este trabajo se centra en la metodología CART la cual se usa para la construcción de árboles de regresión y clasificación, y utiliza un algoritmo recursivo de partición binaria en cada nodo.

II.2. Metodología CART

CART es una técnica exploratoria de datos que tiene como objetivo fundamental encontrar reglas de clasificación y predicción.

Dado un conjunto de datos **D** = (**X**, Y), donde Y es la variable a explicar y **X** = (X_1,. . ., X_p) es un vector de p variables que describe a los individuos, el objetivo de CART es predecir los valores de Y a partir de los valores observados de las variables X_i, i = 1,. . ., p. Tanto la variable dependiente Y, como cada una de las variables explicativas X_i puede ser cuantitativa o cualitativa, esto dota a CART de una gran flexibilidad pues se puede aplicar en muchos contextos distintos.

En el caso en que la variable dependiente Y sea cualitativa, se dice que CART es un **árbol de clasificación**, y el objetivo es predecir la *clasificación* que le correspondería a un individuo con cierto perfil de valores en las variables explicativas. Por otra parte, si Y es cuantitativa, CART es llamado **árbol de regresión** y el objetivo es idéntico al de un modelo lineal, obtener una *estimación* del valor de Y asociado a cada nicho o perfil de predictores.

Además, esta técnica es utilizada para la selección de variables en el sentido que permite determinar cuál característica -o conjunto de características- es la que mejor define o discrimina a los grupos predeterminados.

El Modelo

Los árboles de decisión de tipo CART, pueden verse como la estructura resultante de la partición recursiva del espacio de las variables explicativas (espacio de representación) a partir de un conjunto de reglas de decisión.

La manera en que se construye cada partición es lo que distingue a los distintos tipos de árboles, éstas son determinadas por un conjunto de decisiones sobre las variables explicativas. En CART las reglas de decisión son desplegadas en forma de árbol binario. Determinan en cada momento dos alternativas posibles, las mismas se suceden hasta que el árbol llega a su construcción final. El procedimiento es recursivo y se traduce en una organización jerárquica del

espacio de representación.

La construcción del árbol se basa en tomar, en forma sucesiva distintas decisiones. En cada nodo la decisión es tomada en base a una única variable. La regla de clasificación es sencilla: en cada nodo de decisión se verifica si el valor de cierta variable es mayor que cierto valor específico. Si es mayor se sigue el camino (rama) de la derecha y si es menor el de la izquierda. De este modo, cada decisión da lugar a la partición de los datos en dos grupos mutuamente excluyentes. El nodo inicial es llamado nodo raíz o grupo madre y se divide en dos nodos hijos, luego el procedimiento de partición es aplicado a cada nodo hijo por separado. Las divisiones se seleccionan de modo que la "impureza" de los nodos hijos sea menor que la del nodo madre.

El objetivo es formar grupos homogéneos respecto a la variable que se desea discriminar y a su vez mantener el árbol razonablemente pequeño.

Para dividir los datos se requiere un criterio de particionamiento el cual se basa en una medida de impureza. Esta última establecerá el grado de heterogeneidad de la variable dependiente Y en cada nodo.

El análisis de árboles de clasificación y regresión generalmente consiste en tres pasos (Timofeev, 2004):

1. Construcción del árbol maximal (Construir todas las particiones hasta el final)
2. Poda del árbol (Eliminar las particiones que menos aportan a explicar la respuesta)
3. Selección del árbol óptimo mediante un procedimiento de validación

A continuación se describe cada uno de ellos.

1. Construcción del árbol maximal

El árbol maximal es construido utilizando un procedimiento de partición binario, comenzando en la raíz del árbol. Este árbol es un modelo que describe el conjunto de entrenamiento (grupo de datos original) y generalmente es sobreajustado, es decir, contiene gran cantidad de niveles y nodos que no producen una mejor clasificación, y puede ser demasiado complejo y difícil de interpretar.

Cada nodo es caracterizado por la distribución (cuando Y es cualitativa), o por la media (cuando Y es cuantitativa) de la variable respuesta, el tamaño del nodo y los valores de las variables explicativas que lo definen. Gráficamente el árbol se representa en forma invertida, con el nodo raíz (los datos sin ninguna división) al iniciar, y las ramas y hojas debajo (Figura 1).

Para cada nodo t se tiene un conjunto de reglas de decisión S. Las observaciones que cumplan con la regla van a la izquierda (t_i) y las restantes van a la derecha (t_d).

En cada nodo las particiones son evaluadas. Para seleccionar la regla que provea la "mejor" partición, es necesario tener un criterio de bondad de ajuste de la partición, que se basa en una medida de la impureza de cada nodo, denotada por I(t). La función de impureza es una medida que permite determinar la calidad de un nodo. Está asociada a la heterogeneidad de la variable dependiente Y en ese nodo. Existen varias medidas de impureza que permiten analizar varios tipos de respuesta. Las tres más comunes presentadas por Breiman et al. (1984), para árboles de clasificación son: el índice de información o entropía, el índice de Gini y el índice "Towing".

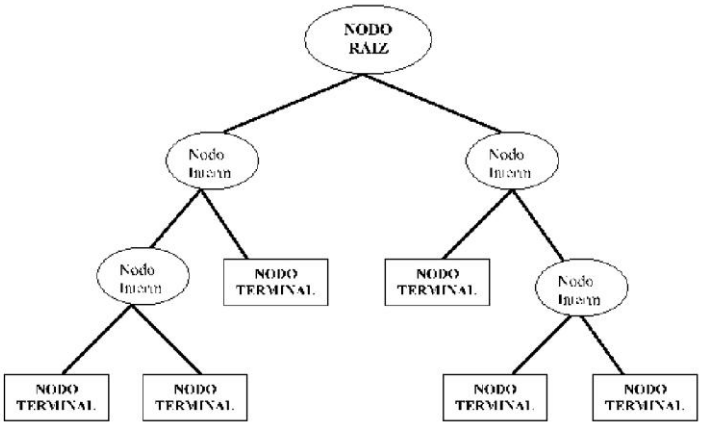

Figura 1: Un árbol genérico

Para cada regla de decisión $s \in S$ se considera $\varphi(s, t) = I(t) - I(t_i) - I(t_d)$, que representa la caída en la función de impureza que se produce al utilizar la regla s para dividir el nodo t. Luego la regla elegida será aquella que maximice $\varphi(s, t)$, es decir $s^* = \max_{s \in S} \{\varphi(s, t)\}$. Cabe aclarar que la maximización se realiza considerando el conjunto S de decisiones, que involucra todas las particiones posibles teniendo en cuenta las p variables explicativas.

2. Poda del árbol

Si las particiones se continúan hasta el final, es posible que los nodos terminales del árbol obtenido sean puros, es decir, que contengan individuos pertenecientes a la misma clase. Sin embargo, este árbol es generalmente sobreajustado. Por este motivo Breiman et al. plantean la necesidad de una vez construído el árbol en forma exhaustiva, proceder a la poda del mismo, cortando sucesivamente ramas o nodos terminales que representen "poco" aporte a la explicación de la variable respuesta, encontrando así el tamaño "adecuado" del árbol.

Breiman et al. (1984) introducen algunas ideas básicas para resolver el problema de seleccionar el mejor árbol. Computacionalmente el procedimiento descripto es complejo. Una forma es buscar una serie de árboles anidados de tamaños decrecientes (De'ath & Fabricius, 2000), cada uno de los cuales es el mejor de todos los árboles de su tamaño.

Estos árboles pequeños son comparados para determinar el óptimo. Esta comparación está basada en una función de Costo-Complejidad, denotada por $R_\alpha(T)$.

Para cada árbol T, la función de Costo-Complejidad se define como (Deconinck et al., 2006):

$$R_\alpha(T) = R(T) + \alpha|\dot{T}| \qquad (1)$$

donde R(T) puede ser el promedio de la suma de cuadrados entre los nodos, la tasa de mala clasificación total o la suma de cuadrados residuales total, dependiendo del tipo de árbol; $|\dot{T}|$ es la complejidad del árbol T, definida como el número total de nodos del sub-árbol y α es el parámetro de complejidad.

El parámetro α es un número real mayor o igual a cero. Cuando $\alpha = 0$ se tiene el árbol más

grande y a medida que α se incrementa, se reduce el tamaño del árbol.

La función $R_\alpha(T)$ siempre será minimizada por el árbol más grande, por tanto se necesitan mejores estimadores del error. Para esto Breiman et al. (1984) propone obtener estimadores "honestos" del error por "validación cruzada".

3. Selección del árbol óptimo

De la secuencia de árboles anidados es necesario seleccionar el árbol óptimo y para esto no es efectivo utilizar comparación o penalización de la complejidad (De'ath & Fabricius, 2000), por tanto se requiere estimar con precisión el error de predicción.

El objetivo es encontrar la proporción óptima entre la tasa de error (tasa de mala clasificación) y la complejidad del árbol, siendo la tasa de error igual al cociente entre el número de individuos mal clasificados y el número total de individuos.

El procedimiento de validación puede implementarse de dos formas:

- Si se cuenta con suficientes datos se divide el conjunto total en dos subconjuntos complementarios, se construye la secuencia de árboles con uno de ellos (denominado datos de entrenamiento o training set), y luego se predice para cada árbol, el valor de la respuesta del otro subconjunto (denominado datos de prueba o test set). Se selecciona el árbol con el menor error de predicción.

  En general no se cuenta con suficientes datos como para utilizar el procedimiento anterior, de modo que otra forma sería:

- Validación cruzada

  Existen diferentes tipos de validaciones cruzadas. Para la metodología CART generalmente se utiliza Validación Cruzada con partición en k (k-fold cross-validation). En particular, cuando k = 1, se tiene el método de Validación Cruzada dejando uno fuera (Leave-one-out cross-validation). La idea básica de este método es que cada individuo es clasificado a partir del árbol construido por todos los individuos de la muestra excepto él. Este proceso se realiza para cada uno de los árboles de la secuencia, y se selecciona aquel que minimice la tasa de error.

III. Aplicación a un problema médico

Descripción del problema

La tuberculosis pulmonar ofrece un modelo atractivo para la investigación de los procesos patológicos que ocurren durante una enfermedad infecciosa. En primer lugar, sigue siendo un problema de salud y una carga económica importante en todo el mundo. En segundo lugar, la existencia de manifestaciones metabólicas relacionadas, ofrece una oportunidad extraordinaria para estudiar la componente inmuno-endócrina que puede ser la base de estas alteraciones, teniendo en cuenta que un suministro de energía razonablemente estable es necesario para preservar todas las funciones biológicas, por ejemplo, la respuesta inmune (Santucci et al, 2011).

Estudios en pacientes con tuberculosis mostraron que una serie de alteraciones inmuno-

endocrinas estaban caracterizadas por distintos niveles de variables hormonales y de proteínas (del Rey, 2007; Mahuad, 2007).

Debido a la necesidad de comprender la respuesta inmune de un modo integrado y del interés en evaluar la relación entre ciertos patrones inmune-endocrino (leptina, adiponectina, IL-6, IL-1b, ghrelina, y cortisol entre otros) y el estado clínico de los pacientes tuberculosos, sus convivientes y controles sanos, se empleará un enfoque que permite analizar varias variables simultáneamente.

Descripción de los datos

Se estudiaron 105 individuos, 53 pacientes con Tuberculosis Pulmonar (TB), 27 convivientes sin diagnóstico previo de la enfermedad y 25 controles hospitalarios sanos.

Los pacientes incluidos fueron casos nuevos de diagnóstico de TB que no presentaban coinfección de VIH/SIDA. El diagnóstico se basó en información clínica y radiológica junto con la identificación del bacilo de la tuberculosis en el examen del esputo.

Los convivientes elegidos eran contactos (VIH-1 seronegativos) de primer orden que compartieron la casa o un cuarto con los pacientes con TB por lo menos tres meses antes del diagnóstico. Fueron evaluados cuidadosamente en base a los exámenes clínicos y radiológicos para descartar tuberculosis.

Los controles hospitalarios sanos, de iguales características socio-económicas, no tenían contacto previo con pacientes con TB ni evidencia clínica o radiológica de la enfermedad.

Tanto los convivientes como los controles sanos no tenían otras enfermedades respiratorias o enfermedades o terapias inmunocomprometedoras.

Se obtuvieron las muestras de sangre para todos los donantes al entrar al estudio, en los pacientes con TB antes de comenzar el tratamiento antituberculoso.

Las variables bajo estudio para caracterizar y poder clasificar a las observaciones fueron la edad, el índice de masa corporal (calculado como el peso sobre la altura al cuadrado, kg/m2) y todas las analizadas de las muestras de sangre: Leptina (pg/ml), Adiponectina (ng/ml), PCR (mg/ml), DHEA (ng/ml), Grhelina (ng/ml), Cortisol (ng/ml), IL-6 (pg/ml), y IL-1b (pg/ml).

El análisis de los datos se realizó con el software estadístico R versión 2.11.1 utilizando la librería “rpart” para la aplicación de CART.

Resultados

Un problema de clasificación real se presenta al estudiar las variables obtenidas en análisis de sangre sobre los tres grupos bajo estudio: pacientes enfermos de tuberculosis pulmonar, convivientes de estos pacientes y controles sanos (Santucci et al, 2011).

Análisis univariado

Inicialmente se lleva a cabo un análisis descriptivo univariado de los grupos mediante medidas resúmenes (mínimo, máximo, media, desviación estándar, mediana y cuartiles). El mismo se presenta en la Tabla 1. Es posible observar notorias diferencias entre los grupos en cuanto a medidas de posición central y variabilidad.

Para analizar dichas diferencias se aplica la prueba no paramétrica de Kruskal Wallis (Tabla 2) utilizando un nivel de significación del 5%. Las variables que presentan diferencias significativas

son: BMI, Leptina, PCR, DHEA, IL-6, IL-1b, Adiponectina y Cortisol.

Tabla 1. Medidas resúmenes para las variables bajo estudio según grupo

Grupo	Variables	Mínimo	Máximo	Media	Desvío estándar	Mediana	Primer cuartil	Tercer cuartil
TB (N= 53)	Edad	15,00	71,00	37,40	16,06	33,00	24,00	50,50
	BMI	16,50	32,00	21,57	2,89	20,90	19,70	22,80
	Leptina	99,00	24724	3540,00	5472,00	1602,00	758,00	3319,00
	Adiponectina	1952,00	23140	10671	4821,00	10256	7068,00	13714
	PCR	3,30	100,00	35,27	24,83	41,90	8,65	49,40
	DHEA	0,82	21,43	4,24	3,20	3,61	2,26	5,32
	Grhelina	105,40	1469,00	460,30	342,20	313,80	196,00	629,60
	Cortisol	40,10	1325,30	186,20	191,50	137,90	105,10	195,90
	IL-6	1,20	53,90	8,46	9,33	6,30	2,10	10,20
	IL-1b	0,20	25,50	0,95	3,48	0,34	0,20	0,42
Convivientes (N=27)	Edad	18,00	72,00	44,26	15,45	46,00	30,00	57,00
	BMI	18,70	48,70	27,42	6,03	27,00	23,80	30,40
	Leptina	765,00	43687	15569	15844	5430,00	2824,00	31030
	Adiponectina	2777,00	29638	9806,00	5990,00	7693,00	5951,00	13422
	PCR	3,30	67,90	9,23	15,19	3,40	3,30	3,95
	DHEA	1,01	18,12	6,18	4,26	4,04	2,77	8,96
	Grhelina	119,20	1060,10	345,20	267,40	229,10	143,30	501,00
	Cortisol	11,13	208,46	107,50	47,44	113,19	66,24	139,78
	IL-6	0,40	57,40	3,78	10,80	1,10	1,10	2,20
	IL-1b	0,20	1,42	0,29	0,24	0,20	0,20	0,26
Controles (N=25)	Edad	20,00	60,00	35,32	13,00	34,00	24,00	45,00
	BMI	18,30	39,10	28,22	4,69	28,40	25,30	31,05
	Leptina	587,00	67093	13375	15898	7005,00	4709,00	16529,00
	Adiponectina	2400,00	17991	7736,00	3997,00	6703,00	4740,00	9515,00
	PCR	3,20	55,51	6,45	10,69	3,30	3,30	3,45
	DHEA	3,06	54,35	9,71	10,52	6,64	4,46	10,07
	Grhelina	143,30	812,80	345,20	203,20	250,90	198,40	479,20
	Cortisol	68,60	240,80	135,80	52,30	123,00	90,40	181,50
	IL-6	0,25	18,30	2,03	3,67	1,00	0,80	1,15
	IL-1b	0,20	2,64	0,31	0,49	0,20	0,20	0,23

Tabla 2. Análisis de las diferencias entre grupos. Test de Kruskal Wallis

Variable	Chi-cuadrado	gl	p-value
Edad	4,80	2	0,091
BMI (kg/m2)	40,03	2	0,000
Leptina	30,48	2	0,000
Adiponectina	7,70	2	0,021
PCR	39,17	2	0,000
DHEA	18,94	2	0,000
Grhelina	3,22	2	0,200
Cortisol	7,00	2	0,030
IL-6 (pg/ml)	48,96	2	0,000
IL-1b (pg/ml)	17,84	2	0,000

Árboles de clasificación y Regresión

El análisis se realiza dividiendo el conjunto total de datos en dos subconjuntos complementarios e independientes entre sí: muestra de entrenamiento y muestra test. Se extrae en forma aleatoria una muestra correspondiente a un tercio de los datos. Estos individuos no intervienen en la construcción del árbol de clasificación, conforman la muestra test, con la cual se realiza luego la validación del mismo.

En la Tabla 3 se presentan los individuos clasificados según muestra a la que pertenecen y grupo de estudio.

Tabla 3. Individuos clasificados según muestra y grupo de estudio

	Muestra	Grupo			Total
		TB	Convivientes	Control	
N	**Entrenamiento**	34	20	16	70
	Test	19	7	9	35
	Total	53	27	25	105
%	**Entrenamiento**	48,57%	28,57%	22,86%	100,00%
	Test	54,29%	20,00%	25,71%	100,00%
	Total	50,48%	25,71%	23,81%	100,00%

Para la construcción del árbol se utiliza como medida de impureza el índice de Gini y probabilidades a priori proporcionales a la frecuencia de cada grupo. Además se decide fijar un número mínimo de 10 observaciones por nodo como condición de parada del proceso recursivo de divi-

siones.

Con el objetivo de encontrar el árbol óptimo, se construye una secuencia de árboles anidados considerando distintos valores para el parámetro de complejidad α. Se busca encontrar aquel árbol que proporcione un equilibrio entre complejidad (tamaño del árbol) y poder predictivo (Tabla 4).

Tabla 4. Búsqueda del árbol óptimo: evaluación Costo-Complejidad

Árbol	α	Tamaño	Tasa de error global	
			Muestra Entrenamiento	Muestra Test
A_1	0	6	15,71%	25,71%
A_2	0,03	5	17,14%	20,00%
A_3	0,06	4	20,00%	25,71%
A_4	0,12	3	25,71%	22,86%
A_5	0,17	2	34,28%	25,71%

En la Tabla 4 es posible observar que el árbol A_1 está sobreajustado a los datos de entrenamiento. Al evaluar su desempeño predictivo con los datos de control o test, se observa que la menor tasa de error se obtiene mediante el segundo árbol de la secuencia. Es decir que aumentar el tamaño de 5 a 6 estaría complejizando el árbol y no mejora la clasificación. Como consecuencia se elige el árbol A_2 (Figura 2).

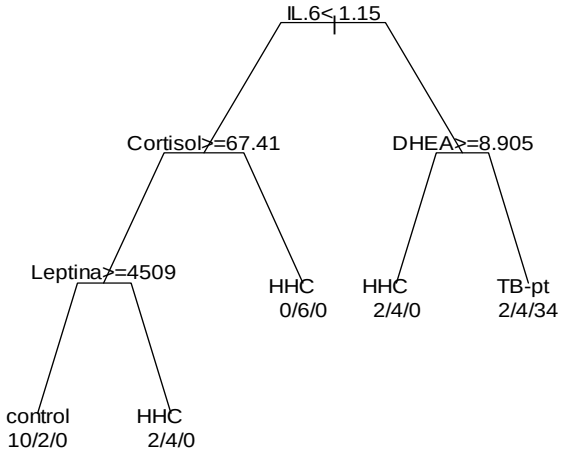

Figura 2: Clasificación de los individuos según el Árbol A_2

Los resultados encontrados permiten caracterizar los siguientes grupos de individuos según clasificación por perfil:

- El perfil con alta probabilidad (85%) de ser clasificado como TB se caracteriza por tener valores de IL-6 mayores o iguales a 1,15 pg/ml y de DHEA inferiores a 8,905 ng/ml.

- El perfil con alta probabilidad (83,33%) de ser clasificado como Control corresponde a los individuos con valores de IL-6 inferiores a 1,15 pg/ml, valores de Cortisol mayores o iguales a 67,41 ng/ml y de Leptina que superan o igualan los 4509 pg/ml.
- El primer grupo de sujetos con alta probabilidad (100%) de ser clasificados como Convivientes son aquellos con valores de IL-6 inferiores a 1,15 pg/nl y de Cortisol menores a 67,41 ng/ml.
- El segundo grupo con alta probabilidad (66,67%) de ser clasificado como Conviviente corresponde a los individuos con valores de IL-6 menores a 1,15 pg/nl, con valores de Cortisol superiores o iguales a 67,41 ng/ml y de Leptina inferiores a 4509 pg/ml.
- El último agrupamiento se caracteriza por tener valores de IL-6 mayores o iguales a 1,15 pg/nl y DHEA no menor a 8,905 ng/ml. Estos individuos tienen alta probabilidad (66,67%) de ser clasificados como Convivientes.

De acuerdo a la estructura jerárquica del Árbol de clasificación elegido (Figura 2), es posible decir que las variables que más diferencian o mejor definen a los grupos en estudio, en orden de importancia, son: IL-6, Cortisol, DHEA y Leptina.

En el proceso de selección del árbol óptimo, se observó que la tasa de error global obtenida al clasificar los individuos test con A_2 es del 20% (Tabla 4). El árbol elegido tiene un buen desempeño tanto a nivel global (80% de acierto), como en forma marginal (para cada uno de los grupos). El porcentaje de clasificación correcta es del 89,5% para los pacientes con Tuberculosis, del 71,4% para los Convivientes y del 66,7% para los Controles sanos (Tabla 5).

Tabla 5. Clasificación de los individuos test según el Árbol A_2

	Grupo observado	Grupo predicho			Total
		TB	Convivientes	Control	
N	TB	17	2	0	19
	Convivientes	2	5	0	7
	Control	2	1	6	9
%	TB	89,47%	10,53%	0,00%	100,00%
	Convivientes	28,57%	71,43%	0,00%	100,00%
	Control	22,22%	11,11%	66,67%	100,00%

De los resultados obtenidos se puede apreciar que la técnica no paramétrica de CART consigue mejorar la clasificación general obtenida mediante las técnicas Análisis Discriminante y Vecinos más cercanos (utilizando k=3 vecinos) (con tasas de acierto del 74,3% y 78% respectivamente). Además permite mantener la buena clasificación para los pacientes con Tuberculosis (aproximadamente del 90%) y mejorar la correcta clasificación tanto para los individuos convivientes como para los controles sanos (Santucci et al., 2011).

IV. Consideraciones finales

La técnica de Árboles de Clasificación y Regresión propuesta como alternativa ante el incumplimiento de los supuestos requeridos por los métodos clásicos como el Análisis Discriminante, es una herramienta sumamente flexible que presenta ventajas sobre otras técnicas. Entre ellas:

- No requiere validar supuestos distribucionales de probabilidad.
- Realiza selección de variables en forma automática, brindando una medida de importancia en forma natural.
- Permite trabajar con variables explicativas y respuesta de tipo cualitativa y/o cuantitativa.
- Es invariante ante transformaciones monótonas de las variables de intervalo.
- Permite valores faltantes para las variables explicativas en los individuos, tanto en la fase de construcción del árbol como en la de predicción.
- Es fácil de interpretar y muy rápida para su implementación.

Sin embargo es importante tener en cuenta la inestabilidad de este método, que es su principal desventaja. Esto significa que pequeñas diferencias en los datos de entrenamiento pueden dar lugar a clasificadores muy distintos. Por este motivo se sospecha que el método de validación utilizado no es el más preciso, debido a la posible variación de los resultados obtenidos para diferentes muestras de entrenamiento. La evaluación podría depender en gran medida de cómo es la división entre datos de entrenamiento y test, y por lo tanto ser significativamente diferente en función de cómo se realice esta división (Schneider, J., 1997; Refaeilzadeh, P., 2008).

Debido a estas carencias se propone evaluar el desempeño de CART mediante validación cruzada y seguir avanzando en el estudio de otras técnicas como Bagging, Random Forest y Vecinos más cercanos que suelen ser métodos más estables.

REFERENCIAS BIBLIOGRÁFICAS

Agresti, A. An Introduction to Categorical Data Analysis. Wiley & Sons, 1996.

Agresti, A. Categorical Data Analysis. Wiley & Sons, 2002.

Ariel Roche, Tesis de Maestría en Ingeniería Matemática Facultad de Ingeniería, UDE-LAR (2009). "Árboles de decisión y Series de tiempo".

Barajas, F. H. (2007). Comparación entre análisis discriminante no-métrico y regresión logística multinomial. Tesis de Maestría, Facultad de Ciencias, Universidad Nacional de Colombia.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. G. (1984). Classification and Regression Trees. Wadsworth International Group, Belmont, California, USA.

Blanco, J. (2006). Introducción al Análisis Multivariado. IESTA. Montevideo.

Castrillón, F. (1998). Comparación de la discriminación normal lineal y cuadrática con la regresión logística para clasificar vectores en dos poblaciones. Tesis de Maestría, Facultad de ciencias, Universidad Nacional de Colombia.

Cuadras, Carlos M. (2010). "Nuevos Métodos de Análisis Multivariante".

De'ath, G. & Fabricius, K. E. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. Ecology, 81 (11), 3178–3192.

Deconinck, E., Zhang, M. H., Coomans, D., & Heyden, Y. V. (2006). Classification tree models for the prediction of blood-brain barrier passage of drugs. Journal of Quemical Information and Modeling, 46 (3), 1410–1419.

del Rey A, Mahuad C, Bozza V, Bogue C, Farroni M, et al. (2007) Endocrine and cytokine responses in humans with pulmonary tuberculosis. Brain Beba Immun 21: 171–179.

Everitt, B. (2005). An R companion to Multivariate Analysis. Springer. London.

Gérard Biau, Luc Devroye, Gábor Lugosi (2008). “Consistency of random forests and other averaging classifiers”.

Hadidi, N. (2003). “Classification ratemaking using decision trees”. CAS Forum.

Härdle & Müller (1997). “Multivariate and Semiparametric Kernel Regression”.

Hechenbichler & Schliep (2004). “Weighted k-Nearest-Neighbor Techniques and Ordinal Classification”. Sonderforschungsbereich 386, Discussion Paper 399.

Hosmer D., Lemeshow S. (2000). “Applied Logistic Regression”. Wiley & Sons.

Johnson, D. (2000) “Métodos Multivariados Aplicados al Análisis de Datos”. International Thompson Editores.

Journal of Artificial Intelligence Research 2 (1994) 1-32. “A System for Induction of Oblique Decision Trees”.

Journal of Computational and Graphical Statistics (2003), 12, 512–530. “Classification Trees with Bivariate Linear Discriminant Node Models”.

Khattree R., Naik D. (2000). Multivariate Data Reduction and Discrimination with SAS® Software. Cary, NC: SAS Institute Inc.

Kovalevski, L. (2011). “Métodos de clasificación no paramétrica”. Decimosextas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística, Universidad Nacional de Rosario, Rosario, Argentina.

Lebart, L., Morineau, A., Piron, M. (1995). Statistique exploratorie multidimensionnelle. Du-nod. Paris.

López, M;et al.(2007) A comparison of classification tree and linear regression analysis for the assessment of vaccine quality. 56th Session-ISI. Book of Abstracts, 274.

Mahuad C, Bozza V, Pezzotto SM, Bay ML, Besedovsky H, et al. (2007) Impaired Immune Responses in tuberculosis patients are related to weight loss that coexists with an immunoendocrine imbalance. Neuroimmunomodulation 14: 193–199.

Manel Martínez Ramón (2008) “Introducción a los métodos Kernel”. Universidad Autónoma de Madrid. 29 de abril de 2008. Universidad Carlos III de Madrid. Departamento de Teoría de la Señal y Comunicaciones.

Nisbet,R.; Elder,J; Miner,G.(2009) Handbook of Statistical Analysis & Data Mining. Elsevier.

Peña, D. (2002). Análisis de datos multivariantes. McGraw-Hill. Madrid.

Peña, D. (2004). Análisis Multivariante. Mc.Graw Hill

Raiko,T.;Ilin,A.;Karhunen,J. (2007). Principal component analysis for large scale problems with lots of missing values. Lecture Notes in Computer Science,4701,691-698. Springer-Verlag.

Refaeilzadeh, P. Tang, L. Lui, H. (2008). K-fold Cross-Validation, Arizona State University.

Samuel Robert Reid, Machine Learning CSCI 5622 (2004). “Decreasing the Randomness of Random Forests”.

Santucci N, D’Attilio L, Kovalevski L, Bozza V, Besedovsky H, et al. (2011). A Multifaceted Analysis of Immune-Endocrine-Metabolic Alterations in Patients with Pulmonary Tuberculo-sis. PLoS ONE 6(10): e26363. doi:10.1371/journal.pone.0026363

Schneider, J. (1997). The holdout method. The school of science.

Skillicorn,D.(2007) Understanding Complex Data Sets. Data Mining with Matrix Descompositions. Chapman and Hall.

Timofeev, R. (2004). Classification and regression trees (cart). theory and applications. Master thesis, CASE - Center of Applied Statistics and Economics. Humboldt University, Berlin.

Torres,P.; Quaglino,M. Pillar,V.(2010) Properties of a randomization test for multifactor comparisons of groups. J.Statistical Computation and Simulation,80,10,1131/50. Londres.

Usuga, O. (2006). Comparación entre análisis de discriminante no-métrico y regresión logística. Proceedings of the Federal American Society of Experimental Biology, 31, 58–61.