Kovalevski, Leandro[1]

[1] *Instituto de Investigaciones Teóricas y Aplicadas, Escuela de Estadística, Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario, Rosario, Argentina*

MÉTODOS DE CLASIFICACIÓN NO PARAMÉTRICA.

I. Introducción

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 manifestaciones metabólicas relacionadas, ofrece una oportunidad extraordinaria para estudiar la componente inmuno-endócrina que puede ser la base 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 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.

El análisis requerido por el problema demanda la utilización de alguna técnica de análisis de datos que permita caracterizar, diferenciar o clasificar a grupos de individuos en base a un conjunto de variables medidas sobre los mismos, que puede ser importante en número. Una de las técnicas más utilizadas para este fin es el Análisis Discriminante. (Khattree,R., Naik,D., 2000; Johnson, D., 2000). La idea básica que persigue este método es determinar si ciertos grupos previamente definidos, difieren entre sí, considerando alguna función lineal o cuadrática de variables, y emplear luego esa función para predecir la pertenencia de una nueva observación a alguno de los grupos. El Análisis Discriminante es 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).

Ante las situaciones donde el Análisis Discriminante no es óptimo se proponen los métodos no paramétricos de clasificación, entre ellos: el método basado en los k vecinos más cercanos, el basado en núcleos Kernel, los árboles de clasificación y regresión (CART: Classification and Regression Trees) y las redes neuronales. (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 una de estas técnicas multivariadas de clasificación, el método basado en los k vecinos más cercanos, en un contexto donde los métodos clásicos no son los más adecuados para el análisis derivado del problema médico expuesto por las características de las variables estudiadas.

II. Material y Métodos

II.1 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/m^2) 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).

II.2 Análisis discriminante

Mediante el Análisis Discriminante se busca extraer a partir de $x_1, x_2, \ldots, x_p$ variables observadas en k grupos, r funciones $y_1, \ldots, y_r$ de forma:

$y_i = a_{i1}x_1 + \cdots + a_{ip}x_p + a_{i0}$

siendo $r = \min(p, k-1)$ y tales que $corr(y_i, y_j) = 0$ para todo $i \neq j$.

Las funciones $y_1, \ldots, y_r$ se construyen de modo que:

- y_1 sea la combinación lineal de $x_1, x_2, \ldots, x_p$ que mejor diferencia a los k grupos.
- y_2 sea la combinación lineal de $x_1, x_2, \ldots, x_p$ que mejor diferencia los k grupos (después de y_1) tal que $corr(y_2, y_1) = 0$
- etc.

II.2.1 Clasificación de los individuos

Conociendo las p variables $x_1, x_2, \ldots, x_p$ sobre un nuevo individuo es posible clasificarlo en uno de los k grupos a partir de las funciones discriminantes $y_1, \ldots, y_r$.

Se calculan los valores para esas r funciones en el individuo nuevo y luego la distancia a cada uno de los vectores de las r funciones valorizadas en los promedios de los k grupos. El nuevo individuo se asigna al grupo cuyo promedio se encuentra a menor distancia.

Si además se conoce la probabilidad a priori de que un individuo pertenezca a cada uno de los k grupos, puede usarse para mejorar la clasificación.

Para que la clasificación a través del Análisis Discriminante sea óptima las variables originales deben seguir una distribución normal multivariada y las matrices de variancia y covariancias deben ser iguales en todos los grupos. Cuando estas condiciones no se cumplen y además hay presencia de outliers se recurre a los métodos de clasificación no paramétricos, entre los que podemos destacar al método de los k vecinos más cercanos.

II.3 Método de Clasificación de los k vecinos más cercanos

Este método de clasificación no paramétrico es conceptualmente simple, consiste en calcular la distancia de un individuo de la muestra a todo el resto, examinar el grupo de pertenencia de las k observaciones más cercanas al individuo y asignar al individuo al grupo con mayor presencia en esas k observaciones.

Para utilizar este método es necesario definir con cuantos vecinos se va a trabajar, es decir, definir el valor de k y elegir una medida con la que ve va a medir la distancia entre las observaciones.

La distancia más conocida que se puede elegir es la euclídea, que está dada por la raiz cuadrada de la suma de las diferencias al cuadrado de cada variable. Es decir, dados dos individuos $\mathbf{x^{(1)}} = (x_1^{(1)}, x_2^{(1)}, \ldots, x_p^{(1)})$ y $\mathbf{x^{(2)}} = (x_1^{(2)}, x_2^{(2)}, \ldots, x_p^{(2)})$, la distancia euclidea entre ellos es:

$$d(x^{(1)},x^{(2)}) = \sqrt{(\boldsymbol{x^{(1)}} - \boldsymbol{x^{(2)}})'(\boldsymbol{x^{(1)}} - \boldsymbol{x^{(2)}})} = \sqrt{(x_1^{(1)} - x_1^{(2)})^2 + (x_2^{(1)} - x_2^{(2)})^2 + \ldots + (x_p^{(1)} - x_p^{(2)})^2}.$$

Esta medida de distancia no es recomendable cuando las variables han sido medidas en unidades muy distintas entre sí. Se pueden estandarizar las variables previamente o utilizar una medida que tenga en cuenta la variabilidad de las variables como la distancia de Mahalanobis, definida por: $d(x^{(1)}, x^{(2)}) = \sqrt{(\boldsymbol{x^{(1)}} - \boldsymbol{x^{(2)}})' \boldsymbol{S}^{-1} (\boldsymbol{x^{(1)}} - \boldsymbol{x^{(2)}})}$ siendo S la matriz de variancia y covariancias.

III. Resultados

Un problema de clasificación real se presentó 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).

III.1 Análisis univariado

Un primer análisis descriptivo de los grupos se presenta en la Tabla 1. Se puede observar diferencias marcadas en las medidas de posición central entre los tres grupos y también grandes diferencias en la variabilidad.

Para probar la significación de esas diferencias entre grupos, se aplicó test no paramétrico de Kruskal-Wallis (Tabla 2). Se observó diferencia (para un nivel de significación del 5%) en las variables: Adiponectina, BMI, Cortisol, DHEA, IL-1b, IL-6, Leptina y PCR.

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

Grupo	Variables	Min	Max	Media	Desv est.	Mediana	P25	P50
TB (N=53)	Edad	15	71	37,40	16,06	34,00	24,00	52,25
	BMI	16,5	32	21,57	2,89	20,90	19,73	22,19
	Leptina	99	24724	3540	5472	1636	791	3173
	Adiponectina	1952	23140	10671	4821	9994	6729	12923
	PCR	3,3	100	35,27	24,83	26,50	7,10	49,41
	DHEA	0,82	21,43	4,24	3,20	4,22	2,37	5,78
	Grhelina	105,4	1469,0	460,3	342,2	296,0	195,6	595,0
	Cortisol	40,1	1325,3	186,2	191,5	127,5	91,4	223,0
	IL-6	1,2	53,88	8,46	9,33	6,34	2,09	10,23
	IL-1b	0,2	25,5	0,95	3,48	0,25	0,20	0,46
Convivientes (N=27)	Edad	18	72	44,26	15,45	46,00	30,00	57,00
	BMI	18,7	48,7	27,42	6,03	27,00	23,80	30,40
	Leptina	765	43687	15569	15844	5430	2824	31030
	Adiponectina	2777	29638	9806	5990	7693	5951	13422
	PCR	3,3	67,93	9,24	15,20	3,35	3,30	3,95
	DHEA	1,01	18,12	6,18	4,26	4,04	2,77	8,96
	Grhelina	119,2	1060,1	345,2	267,4	229,1	143,3	501,0
	Cortisol	11,1	208,5	107,5	47,4	113,2	66,2	139,8
	IL-6	0,35	57,41	3,77	10,81	1,12	1,09	2,17
	IL-1b	0,2	1,42	0,29	0,24	0,20	0,20	0,26
Controles (N=25)	Edad	20	60	35,32	13,00	34,00	25,50	46,00
	BMI	18,3	39,1	28,22	4,69	28,40	24,53	30,70
	Leptina	587	67093	13375	15898	5857	2334	8224
	Adiponectina	2400	17991	7736	3997	6408	4731	8595
	PCR	3,2	55,51	6,45	10,69	3,30	3,30	6,28
	DHEA	3,06	54,35	9,71	10,52	7,90	4,42	12,18
	Grhelina	143,3	812,8	345,2	203,2	288,0	209,7	567,2
	Cortisol	68,6	240,8	135,8	52,3	145,9	113,6	190,4
	IL-6	0,25	18,3	2,03	3,67	0,99	0,66	2,50
	IL-1b	0,2	2,64	0,31	0,49	0,20	0,20	0,21

Tabla 2. Prueba de Kruskal-Wallis para la diferencia entre grupos

Variable	**Chi-Square**	**gl**	**p-value**
Edad	4,799	2	0,091
BMI (kg/m2)	39,994	2	0,000
Leptina	30,479	2	0,000
Adiponectina	7,699	2	0,021
PCR	39,183	2	0,000
DHEA	18,944	2	0,000
Grhelina	3,220	2	0,200
Cortisol	6,997	2	0,030
IL-6 (pg/ml)	48,561	2	0,000
IL-1b (pg/ml)	17,873	2	0,000

III.2 Análisis discriminante

Al analizar cuales son las variables que más diferencian a los grupos a través de un Análisis Discriminante se observó que BMI, PCR y DHEA son las variables que más pesan en la primera función discriminante mientras que Edad, Leptina, DHEA y Cortisol son las que más influyen en la segunda función discriminante (Tabla 3).

Tabla 3. Coeficientes estandarizados de las funciones dos discriminantes

	Función Discriminante	
	1	2
Edad	,088	**,639**
BMI	**,557**	-,259
Leptina	,170	**,389**
Adiponectina	-,028	,239
PCR	**-,509**	-,104
DHEA	**,323**	**-,366**
Grhelina	-,023	,019
Cortisol	-,093	**-,335**
IL-6	-,201	,065
IL-1b	-,090	-,121

Al utilizar las funciones discriminantes para clasificar los individuos, se logran clasificar correctamente a 78 de los 105 individuos (74,3%) si utilizamos todos los individuos mientras que si clasificamos los individuos por validación cruzada, es decir que cada individuo es clasificados por funciones construidas a partir de todos los casos excepto él, el porcentaje de correcta clasificación baja a 63,8% (Tabla 4).

Se puede observar también que el porcentaje de correcta clasificación es alto para los pacientes con TB (92,5% y 84,9% usando todos los datos y por validación cruzada, respectivamente) mientras que para los otros grupos el porcentaje correctamente clasificado es considerablemente inferior.

Tabla 4. Clasificación a través del Análisis Discriminante Paramétrico

		Grupo	Grupo Predicho			Total
			Controles	Convivientes	TB	
Original	N	Controles	16	6	3	25
		Convivientes	7	13	7	27
		TB	0	4	49	53
	%	Controles	64%	24%	12%	100%
		Convivientes	25,9%	48,1%	25,9%	100%
		TB	0%	7,5%	92,5%	100%
Validación cruzada	N	Controles	13	9	3	25
		Convivientes	8	9	10	27
		TB	1	7	45	53
	%	Controles	52%	36%	12%	100%
		Convivientes	29,6%	33,3%	37%	100%
		TB	1,9%	13,2%	84,9%	100%

III.3 Método de clasificación de los k vecinos más cercanos.

Con este método no paramétrico de clasificación no se crea un modelo solo debe definirse el número de vecinos a utilizar y le medida de la distancia.

En este caso se decidió trabajar con k=3 vecinos y con las distancia de Mahalanobis para considerar las distinta variabilidad de las variables estudias.

Al analizar los resultados de la clasificación de este método no paramétrico (Tabla 5) observamos que se clasifican correctamente 82 de los 105 individuos (78%) utilizando en conjunto de datos completo mientras que si se realiza una clasificación cruzada se clasifican correctamente 72 (69%).

Tabla 5. Clasificación a través del método no paramétrico de los vecinos más cercanos (k=3, es decir tomando los tres vecinos más cercanos)

		Grupo	Grupo Predicho			No definido	Total
			Controles	Convivientes	TB		
Original	N	Controles	20	1	1	3	25
		Convivientes	6	14	4	3	27
		TB	0	2	48	3	53
	%	Controles	80%	4%	4%	12%	100%
		Convivientes	22,2%	51,8%	14,8%	11,1%	100%
		TB	0%	3,75	90,5%	5,6%	100%
Validación cruzada	N	Controles	15	7	3	-	25
		Convivientes	10	9	8	-	27
		TB	0	5	48	-	53
	%	Controles	60%	28%	12	-	100%
		Convivientes	37%	33,3%	29,6%	-	100%
		TB	0%	9,4%	90,5%	-	100%

De los resultados obtenidos se puede apreciar que este método no paramétrico simple propuesto mejora la clasificación general que se había realizado a través del Análisis Discriminante manteniendo la buena clasificación dentro de los pacientes con TB y mejorando la correcta clasificación de los controles sanos.

IV. Consideraciones finales

El método de los k vecinos más cercanos es una alternativa válida para clasificar cuando las condiciones óptimas para el Análisis Discriminante no se presentan.

Se propone seguir avanzando en el estudio de otros métodos de clasificación no paramétricos (el método basado en núcleos Kernel, los árboles de clasificación y regresión y las redes neuronales) y comparar sus resultados con los obtenidos con el método de los k vecinos así como también buscar una estrategia para definir el número k de vecinos a utilizar.

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, UDELAR (2009). "Árboles de decisión y Series de tiempo".

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

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

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".

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 D. Applied Logistic Regression. Wiley & Sons, 2000.

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.

Lebart, L., Morineau, A., Piron, M. (1995). Statistique exploratorie multidimensionnelle. Dunod. 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 immuno-endocrine 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.

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 Tuberculosis. PLoS ONE 6(10): e26363. doi:10.1371/journal.pone.0026363

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

Torrres,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.