Diagnóstico automático en Medicina

por JOSE-MIGUEL BERNARDO
Universidad de Valencia

Artículo publicado en la revista "Estadística Española", del Instituto Nacional de Estadística número 108 correspondiente a Julio - Septiembre 1985

MADRID 1986

ESTADÍSTICA ESPAÑOLA
núm 108 1985, págs 19 a 34

Diagnóstico automático en Medicina

por JOSE-MIGUEL BERNARDO
Universidad de Valencia

RESUMEN

En este trabajo se propone una forma de utilizar el conjunto de atributos que describen las características de un paciente, junto a los signos y síntomas de su enfermedad, para determinar las probabilidades asociadas a cada una de las afecciones que pueden haber causado su estado actual, en base a la información proporcionada por un adecuado banco de datos. El modelo matemático descrito recoge la notable interdependencia que típicamente existe entre los distintos atributos y se sitúa en el marco de la metodología estadística bayesiana

Palabras clave Análisis discriminante; clasificación probabilística; distribución diagnóstica; distribuciones de referencia; métodos Bayesianos; regresión logística.

1 DESCRIPCION DEL PROBLEMA

Sea $\{\delta_1, \ldots, \delta_k\}$ una lista *exhaustiva* de las distintas patologías que son aceptadas como posibles causas de un determinado síndrome. Sin pérdida de generalidad, supondremos que describen patologías mutuamente excluyentes; en efecto, bastaría considerar como desórdenes distintos, las asociaciones de dos o más patologías que se consideren compatibles con el síndrome estudiado para, formalmente, reducir el problema a ese caso

Supongamos que se dispone de un *banco de datos* D constituido por la información proporcionada por n pacientes que presentaron el síndrome objeto de estudio. Específicamente supondremos que, para cada uno de esos pacientes, se conocen los *atributos* que pueden resultar *relevantes* tanto características personales (por ejemplo, edad, sexo, ..) como signos y síntomas (por ejemplo, composición de la orina, electrocardiograma,..) de su estado físico, en el momento en que presentaban el síndrome que se pretende estudiar; supondremos además que la evolución de la enfermedad ha permitido determinar con certeza (por ejemplo, mediante intervención quirúrgica, criptopatología, necropsia,..) la verdadera causa del síndrome en cada uno de los pacientes incluídos en el banco de datos. Así, el banco de datos $D = \{ (Z_j, \delta_{i(j)}) , j = 1, \ldots, n \}$ está constituído por n pares $(Z_j, \delta_{i(j)})$ donde $Z_j = \{z_{1j}, \ldots, z_{sj}\}$ representa el vector columna definido por todos los *atributos* (digamos s) del paciente j (incluyendo tanto sus características personales como las magnitudes que describían su estado físico) y donde $\delta_{i(j)} \in \{ \delta_1, \ldots, \delta_k \}$ es el desorden (enfermedad, patología) concreto que motivó el síndrome al paciente j.

Nuestro objetivo es construir un algoritmo que, para cualquier nuevo paciente que represente el síndrome estudiado, permita determinar las probabilidades asociadas a cada una de sus posibles causas, dados los atributos concretos de ese paciente, y dada la información contenida en el banco de datos. Formalmente, nos proponemos determinar el conjunto de probabilidades

$$\{ p (\delta_i | Z, D, H) , \quad i = 1, \ldots, k \}, \tag{1}$$

donde $p (\delta_i | Z, D, H)$ es la probabilidad de que δ_i sea realmente la causa del síndrome, en el caso de un paciente con atributos Z, teniendo en cuenta la información proporcionada por el banco de datos D y las hipótesis estadísticas H que puedan hacerse sobre su comportamiento. Naturalmente, puesto que hemos supuesto que $\{ \delta_1, \ldots, \delta_k \}$ constituyen una lista exhaustiva de las patologías mutuamente excluyentes que pueden causar el síndrome, el conjunto de probabilidades $\{ p (\delta_i | Z, D, H) , i = 1, \ldots, k \}$ constituye una distribución de (la unidad de) probabilidad, que denominaremos *distribución diagnóstica*.

Ejemplo. Aitchison & Dunsmore (1975, Cap. 11) describen un banco de datos apropiado para el diagnóstico diferencial del *síndrome de Cushing*: un conjunto de problemas poco frecuentes, asociado a la hipersecreción de cortisol por la corteza adrenal. Simplificando el problema, se consideran solamente tres posibles causas del síndrome δ_1 = adenoma, δ_2 = hiperplasia y δ_3 = carcinoma, y se suponen muy improbables sus combinaciones. Se trata de investigar la posibilidad de identificar la causa del síndrome mediante la observación de dos atributos que se consideran relacionados con el problema, los logaritmos naturales de la excreción urinaria (en mg/24 h) de dos metabolitos

esteroides, la tetrahidrocortisona (Z_1) y el pregnanetriol (Z_2), determinados por cromatografía en papel. Se dispone del banco de datos reproducido en la Tabla 1, formado por 21 pacientes de los que se conocen tanto los valores de estos atributos como la verdadera causa del síndrome.

j	Z_1	Z_2	δ
1	1 131	2 460	1
2	1 099	262	1
3	.642	0.000	1
4	1.335	−3 219	1
5	1 411	095	1
6	.642	−.916	1
7	2 116	0.000	2
8	1 335	−1.609	2
9	1 361	−.511	2
10	2.054	.182	2
11	2 208	−.511	2
12	2 734	1.281	2
13	2.041	470	2
14	1 872	−.916	2
15	1 740	−.916	2
16	2 610	−.511	2
17	2 322	1 856	3
18	2 219	2 067	3
19	2.262	1 131	3
20	3 985	916	3
21	2 760	2.028	3

Tabla 1: *Banco de Datos* D

El problema es determinar, con esta información, cuál es la distribución de probabilidad, entre los tres posibles orígenes del síndrome, que corresponde a nuevos pacientes con el síndrome de Cushing de los que se conocen sus excreciones urinarias; por ejemplo, las correspondientes a los cuatro nuevos casos recogidos en la Tabla 2.

j	Z_1	Z_2
22	1 629	−.916
23	2 557	1 609
24	2 565	−.223
25	956	−2 303

Tabla 2: *Nuevos pacientes*

2. LOS MODELOS CONOCIDOS

Las soluciones propuestas en la literatura especializada del problema descrito son variantes de tres metodologías muy distintas entre sí, que describiremos brevemente

2.1. *Análisis discriminante lineal*

En un trabajo pionero, Fisher (1936) se propuso resolver un problema más sencillo; supuso que la mayor parte de la información que permite discriminar entre dos conjuntos de vectores puede resumirse en el valor de una combinación lineal de sus elementos adecuadamente elegida, y demostró que la función lineal de los atributos $\lambda' Z_j = \lambda_1 Z_{1j} + \ldots + \lambda_s Z_{sj}$ que mejor *separa* dos clases δ_1 y δ_2, en el sentido de maximizar la distancia tipificada entre las medias de los valores de $\lambda' Z$ que se obtienen para cada una de las clases, resulta ser la definida por los coeficientes

$$\lambda' = (\lambda_1, \ldots, \lambda_s) = (\bar{Z}_1 - \bar{Z}_2)' \, S^{-1} \tag{2}$$

donde Z_1, Z_2 son los vectores media de los atributos en cada una de las clases y S su matriz combinada de varianzas – covarianzas. En realidad, puede demostrarse (Goel, 1983) que la expresión (2) es la función lineal que mejor separa las dos clases para un conjunto muy amplio de posibles definiciones de distancia.

El concepto de función discriminante lineal puede ser generalizado de distintas maneras al caso en que existen más de dos clases. Una de las más sencillas consiste en considerar el *conjunto* de funciones lineales

$$\lambda_i' Z = (\bar{Z}_i - \bar{Z}_k)' \, S^{-1} Z \,, \qquad i = 1, \ldots, k-1 \tag{3}$$

donde k es el número de clases, $\bar{Z}_i$ es el vector media correspondiente a los, digamos, n_i pacientes que sufren el desorden δ_i, $\bar{Z}_k$ es el vector media correspondiente a los n_k pacientes que sufren el desorden δ_k (arbitrariamente tomado como referencia, frecuentemente el grupo control), y S es la matriz combinada de varianzas–covarianzas

$$S = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Z_j - \bar{Z}_i)(Z_j - \bar{Z}_i)'$$

La obtención de la función discriminante permite una aproximación intuitiva al problema que nos hemos propuesto; en efecto, si el vector $\lambda' Z = (\lambda_1' Z, \ldots, \lambda_{k-1}' Z)$

correspondiente a un paciente con atributos Z, se sitúa muy cerca de la media m_i de los vectores $\{ \lambda' Z_j \,, j = 1, \ldots, n_i \}$ que corresponden a los n_i pacientes con el desorden δ_i, podemos suponer que la probabilidad $p(\delta_i \mid Z, D, H)$ de que ese paciente tenga el síndrome estudiado como consecuencia del desorden δ_i, sea muy alta

Sin embargo, sin hacer nuevas hipótesis, las probabilidades requeridas $\{ p(\delta_i \mid Z, D, H) \,, i = 1, \ldots, k \}$ no pueden ser deducidas de los valores $\lambda' Z$ y $\{ \lambda' Z_j \,, j = 1, \ldots, n \}$.

Ejemplo (cont.)

En el ejemplo descrito en el apartado anterior, las funciones discriminantes resultan ser

$$\lambda_1' Z = 7.1914 \, Z_1 + 1.2654 \, Z_2$$

$$\lambda_2' Z = 2.8077 \, Z_1 + 1.5291 \, Z_2$$

En la Figura 1 se representan mediante un dígito que identifica la enfermedad que sufren, los valores de $\lambda_1' Z$ y $\lambda_2' Z$ correspondientes a los pacientes que constituyen el banco de datos descrito en la Tabla 1; además de una fuerte correlación entre ambos valores, puede observarse cómo los valores correspondientes a una misma enfermedad tienden a agruparse. Los valores correspondientes a los cuatro nuevos pacientes de la Tabla 2, señalados con ● sugieren que se trata de dos casos de hiperplesia, un adenoma y un carcinoma.

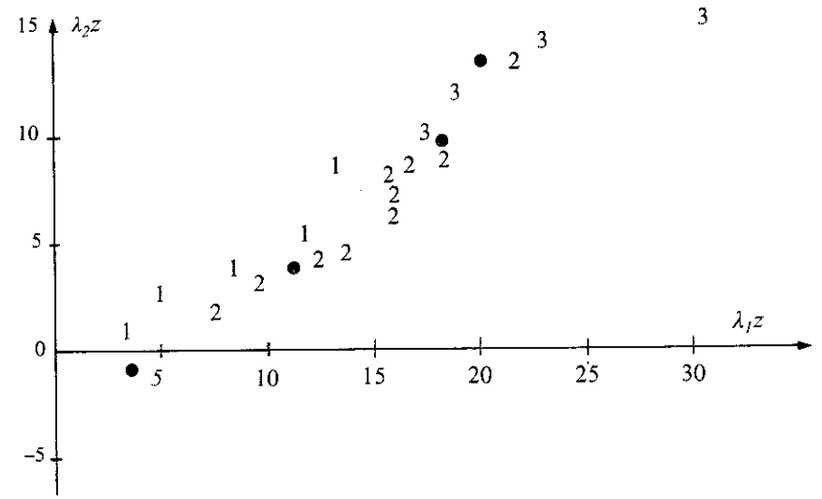

Fig. 1 *Vectores discriminantes de los pacientes contenidos en el banco de datos y de los nuevos casos*

2.2. *El paradigma de muestro*

Una forma de poder *deducir* las probabilidades diagnósticas a partir del banco de datos consiste en suponer que la distribución *conjunta* de los atributos pertenece a una determinada familia para cada uno de los posibles desórdenes δ_i, de forma que $p\ (Z, \delta_i) = p\ (\delta_i \mid \omega, \delta_i)\ p\ (\delta_i)$, donde las distribuciones multivariantes $\{ p\ (\delta_i \mid \omega, \delta_i)\ , i = 1, \ldots, k \}$ son conocidas excepto en el valor concreto del vector paramétrico ω

Si una hipótesis de este tipo resulta razonable, la solución del problema es inmediata. En efecto, los teoremas de Bayes y de la probabilidad total permiten escribir la distribución diagnóstica como

$$p\ (\delta_i \mid Z, D) \propto p\ (Z \mid D, \delta_i)\ p\ (\delta_i \mid D)$$

$$P\ (Z \mid D, \delta_i) = \int p\ (Z \mid \omega, \delta_i)\ p\ (\omega \mid D, \delta_i)\ d\omega$$

$$P\ (\omega \mid D, \delta_i) \propto \prod_{j=1}^{n_i} p\ (Z \mid \omega, \delta_i)\ p\ (\omega \mid \delta_i)$$

donde $p\ (\omega \mid \delta_i)$ es la distribución inicial de referencia (Bernardo, 1979) correspondiente al modelo $p\ (Z \mid \omega, \delta_i)$ y $\{ p\ (\delta_i \mid D), i = 1, \ldots, k \}$ es la distribución de probabilidad sobre los posibles desórdenes que recoge tanto la información inicial como la que pueda haber proporcionado el propio banco de datos

En la práctica, las posibles aplicaciones de este modelo se restringen al caso en el que el conjunto de *todos* los atributos tiene una distribución normal multivariante El análisis correspondiente al caso normal fue originalmente descrito por Geisser (1964) y posteriormente recogido y ampliado por distintos autores (Aitchison and Dunsmore, 1975; Bernardo, 1978; Bermudez, 1979).

Lamentablemente, la hipótesis de normalidad es pocas veces utilizable en las aplicaciones médicas; por ejemplo, si junto a los atributos de tipo cuantitativo, como los escritos en el ejemplo comentado, aparecen atributos cualitativos (por ejemplo, presencia o no de hemorragia, ..) la hipótesis de multinormalidad resulta obviamente inaceptable.

Es muy importante subrayar que la distribución diagnóstica que muchos programas de análisis discriminante producen, es tan solo *la aproximación al resultado exacto del caso multinormal* que resulta de sustituir los parámetros desconocidos por sus estimadores máximo verosímiles En efecto, puede suponerse que

$$p\ (\delta_i \mid Z, D) \propto p\ (Z \mid D, \delta_i)\ p\ (\delta_i \mid D) \simeq p\ (Z \mid \hat{\omega}, \delta_i)\ p\ (\delta_i \mid D)$$

lo que, en el caso normal homocedástico, (Wald, 1945; Anderson, 1951) resulta ser una función de una de las posibles generalizaciones multivariantes de la función discriminante. Específicamente,

$$\log \frac{p\ (\delta_i \mid Z, D)}{p\ (\delta_k \mid Z, D)} = \log \frac{p\ (\delta_i \mid D)}{p\ (\delta_k \mid D)} + \lambda_{io} + \lambda_i' Z\ , \qquad i = 1, \ldots, k-1$$

donde

$$\lambda_{io} = -\frac{1}{2} (\bar{Z}_i - \bar{Z}_k)'\ S^{-1}\ (\bar{Z}_i + \bar{Z}_k)$$

$\lambda_i' = (\bar{Z}_i - \bar{Z}_k)'\ S^{-1}$, ya definido en (3).

Sin embargo, como Aitchison y Dunsmore (1975, p. 228) ponen de manifiesto, con bancos de datos pequeños esta aproximación puede resultar peligrosa, sugiriendo distribuciones diagnósticas mucho más extremas de lo que los datos permiten deducir

2.3. *El paradigma diagnóstico*

Una forma de eludir la especificación de la distribución conjunta de los atributos para cada una de las clases consideradas, consiste en hacer directamente una hipótesis sobre la forma de la distribución diagnóstica (Dawid, 1976). Así, si suponemos que la distribución conjunta de atributos y desórdenes es de la forma

$$p\ (Z, \delta_i) = p\ (\delta_i \mid Z, \theta)\ p\ (Z)$$

donde θ es un parámetro desconocido, y podemos determinar la distribución inicial de referencia correspondiente $p(\theta)$, los teoremas de Bayes y de la probabilidad total permiten escribir la distribución diagnóstica como

$$p\ (\delta_i \mid Z, D) = \int p\ (\delta_i \mid Z, \theta)\ p\ (\theta \mid D)\ d\theta$$

$$p\ (\theta \mid D) \propto p\ (D \mid \theta)\ p\ (\theta)$$

donde $p\ (D \mid \theta)$ es la función de verosimilitud de los datos.

Entre los distintos modelos diagnósticos propuestos, el modelo *logistico aditivo*, definido por las ecuaciones

$$log \frac{p(\delta_i \mid Z, \theta)}{p(\delta_k \mid Z, \theta)} = \theta_{oi} + \theta_i' Z, \qquad i = 1, \ldots, k-1$$

es uno de los más utilizados.

El problema es complejo dada la dificultad que reviste la obtención de la distribución de referencia. Bermudez (1984) ha obtenido resultados interesantes para el caso de muestreo prospectivo; el caso, mucho más frecuente, de datos retrospectivos es notablemente más difícil desde esta perspectiva, debido a la compleja estructura que en este caso adopta la función de verosimilitud. En el caso de datos prospectivos, la llamada *regresión logística* (Cox. 1970, Anderson, 1972) es la aproximación que se obtiene cuando se sustituye al vector paramétrico desconocido θ por su estimador máximo–verosímil; como en el caso del paradigma de muestreo, esta aproximación puede resultar extremadamente peligrosa cuando se trabaja con pequeños bancos de datos. El uso de la regresión logística con datos retrospectivos no resulta justificable.

3 EL MODELO PROPUESTO

El modelo que proponemos en este trabajo constituye una síntesis de los modelos descritos en la sección anterior, que pretende superar las dificultades asociadas a ellos.

Intuitivamente, se utiliza el poder separador de las funciones discriminantes creando un modelo, —dentro del paradígma de muestreo—, que permite deducir de ellas la distribución diagnóstica sin reducir perceptiblemente su campo de aplicabilidad; además, se propone una aproximación al resultado analítico así obtenido que recoge la facilidad de uso y el contenido intuitivo comunmente asociado a los modelos logísticos.

3.1. *Suficiencia de las funciones discriminantes*

Los resultados teóricos ya mencionados sobre la optimalidad de las funciones discriminantes junto a los notables resultados empíricos obtenidos con su uso (ver por ejemplo Press y Wilson, 1978; Titterington *et al.*, 1981) sugieren que, efectivamente, las funciones discriminantes recogen la mayor parte de la información que el banco de datos puede proporcionar para diagnosticar a un nuevo paciente; consecuentemente, supondremos que la distribución diagnóstica sólo depende del vector de atributos a través del valor de sus funciones discriminantes. Formalmente,

$$H1 : p(\delta_i \mid Z, D) = p(\delta_i \mid \lambda' Z, D)$$

donde $\lambda' Z = (\lambda_1' Z, \ldots, \lambda_{k-1}' Z)$ ya fue definido en (3)

3.2 *Distribución del vector discriminante*

Si el número de atributos *no es muy pequeño,* lo que resulta ser la situación habitual en Medicina, pueden invocarse distintas formas, teorema central del límite (por ejemplo Diaconis y Freedman, 1984) para asegurar que la distribución del vector discriminante entre los pacientes con cada uno de los desórdenes será aproximadamente normal, La hipótesis de normalidad también resultará aplicable en situaciones, como la descrita en el ejemplo comentado, en las que se trabaja con un número reducido de *atributos cuantitativos*, apropiadamente transformados si es necesario; en consecuencia, podemos esperar que sea frecuentemente razonable suponer la multinormalidad del vector discriminante. Formalmente,

$$H2 : p(\lambda' Z \mid \delta_i) = N_{k-1}(\lambda' Z \mid \mu_i, \Sigma_i), \qquad i = 1, \ldots, k$$

donde μ_i y Σ_i son parámetros desconocidos.

Naturalmente, esta hipótesis deberá ser contrastada en cualquier aplicación concreta. Específicamente, puede no resultar apropiada cuando se trabaja con un número muy pequeño de atributos que incluye datos cualitativos. En estos casos será necesario modificar H2, sustituyendo la distribución normal por otra más apropiada al problema. Obsérvese sin embargo que, en cualquier caso, el problema planteado es incomparablemente más sencillo que el original, puesto que la dimensión de la distribución de $\lambda' Z$ es $k - 1$ (y muchas veces, especialmente cuando se trabaja en forma secuencial, $k = 2$), mientras que la dimensión de la distribución de Z es s, el número de atributos, frecuentemente superior a 50.

3.3. *Distribución inicial de referencia*

La información inicial sobre los parámetros de que depende la distribución del vector discriminante es típicamente muy difusa, por lo que puede ser aproximada por la correspondiente *distribución de referencia* (Bernardo, 1979 b) que, para un modelo multinormal $N_p(X \mid \mu, \Sigma)$ resulta ser

$$p(\mu, \Sigma) \propto |\Sigma|^{-(p+1)/2}$$

Formalmente, suponemos pues,

$$H3 : p(\mu_i, \Sigma_i \mid \delta_i) \propto |\Sigma_i|^{-k/2}, \qquad i = 1, \ldots, k$$

Naturalmente, si en H2 se propone un modelo distinto al normal multivariante, será necesario modificar H3, sustituyendo la distribución propuesta por la distribución de referencia correspondiente al modelo escogido

3.4 *La distribución diagnóstica*

Con las hipótesis anteriores, es posible *deducir* la distribución diagnóstica. En efecto, por H1 y el Teorema de Bayes

$$p(\delta_i \mid Z, D) = p(\delta_i \mid \lambda' Z, D) \propto p(\lambda' Z) \mid \delta_i, D)\, p(\delta_i \mid D) \tag{4}$$

El primer factor puede ser calculado haciendo uso de los teoremas de Bayes y de la probabilidad total. En efecto,

$$p(\lambda' Z \mid \delta_i, D) = \int\int N_{k-1}(\lambda' Z \mid \mu_i, \Sigma_i)\, p(\mu_i, \Sigma_i \mid D)\, d\mu_i d\Sigma_i$$

donde

$$p(\mu_i, \Sigma_i \mid D) \propto \prod_{j=1}^{n_i} N(\lambda' Z_j \mid \mu_i, \Sigma_i) \mid \Sigma_i \mid^{-k/2};$$

el resultado de esa integral es (Geisser, 1964), la densidad de Student

$$p(\lambda' Z \mid \delta_i, D) = St_{k-1}(\lambda' Z \mid m_i, V_i, n_i - k + 1) \tag{5}$$

donde m_i, V_i son, respectivamente, el vector media y la matriz de covarianzas muestrales de los vectores discriminantes $\{ \lambda' Z_j, j = 1, \ldots, n_i \}$ correspondientes a los pacientes incluídos en el banco de datos que tienen el desorden δ_i.

El segundo factor depende del tipo de muestreo realizado para obtener el banco de datos:

i) Si se trata de un muestreo aleatorio de toda la población estudiada *(datos prospectivos)* la probabilidad asociada a cada categoría resulta ser (DeGroot, 1970)

$$p(\delta_i \mid D) = (n_i + \alpha_i) / (n + \Sigma\alpha_i), \qquad i = 1, \ldots, k \tag{6}$$

donde las α_i describen la información de que se dispone sobre la prevalencia de los distintos desórdenes; si se carece de esta información, debe utilizarse la distribución de referencia, que corresponde a los valores

$$\alpha_i = 1/2, \qquad i = 1, \ldots, k$$

ii) Si se trata de un muestreo aleatorio dentro de cada uno de los desórdenes considerados *(datos retrospectivos)*, el banco de datos no proporciona información sobre la prevalencia de los distintos desórdenes, de forma que,

$$p(\delta_i \mid D) = p(\delta_i), \qquad i = 1, \ldots, k \tag{7}$$

donde las $p(\delta_i)$ describen la información de que se dispone sobre la prevalencia de los distintos desórdenes; de nuevo, si se carece de esta información debe utilizarse la distribución de referencia que, en este caso, es la distribución uniforme

$$p(\delta_i) = 1/k, \qquad i = 1, \ldots, k$$

En resumen, la distribución diagnóstica que se deduce de nuestras tres hipótesis resulta ser

$$p(\delta_i \mid Z, D, H) \propto \mid V_i \mid^{-1/2} \{1 + \frac{1}{n+1} (\lambda' Z - m_i)' V_i^{-1} (\lambda' Z - m_i)\}^{-n_i/2} p(\delta \mid D) \tag{8}$$

donde C_i es la constante de proporcionalidad correspondiente, esto es

$$\{\Gamma(n_i/2)/\Gamma((n_i - k + 1)/2)\}\{\pi(n_i - k + 1)\}^{-(k-1)/2}$$

(que resulta irrelevante si todos los n_i coinciden) y donde $p(\delta_i \mid D)$ viene dado por (6) o por (7) según el tipo de muestreo utilizado para obtener el banco de datos.

3.5 *La aproximación logística*

Aunque la expresión (8) proporciona una solución analítica al problema propuesto con lo que los valores $\{ (n_i, m_i, V_i), i = 1, \ldots, k \}$ y $\{ \lambda_i = (\bar{Z}_i - \bar{Z}_k)' S^{-1}, i = 1, \ldots, k-1 \}$ forman un estadístico suficiente para calcular la distribución diagnóstica correspondiente a cualquier nuevo paciente, para utilizarla es necesario disponer —al menos— de una buena calculadora programable. En consecuencia, para facilitar el uso de la solución propuesta, buscamos una buena *aproximación* que exija poco cálculo.

Ya hemos mencionado en la sección 2.3. que las distribuciones logísticas allí definidas constituyen una amplia familia con la que puede intentar aproximarse cualquier distribución diagnóstica. Desde un punto de vista teórico, se trata de un problema de decisión en el que el espacio de acciones es el conjunto de los posibles valores del parámetro θ y la función de pérdida una medida de la discrepancia entre la verdadera distribución diagnóstica $\{ p(\delta_i \mid Z, D), i = 1, \ldots, k \}$ y su aproximación logística $\{ p(\delta_i \mid Z, \theta), i = 1, \ldots, k \}$

Bajo condiciones muy generales (Bernardo, 1979*a*, 1980) la medida de discrepancia más adecuada es la divergencia dirigida (Kullback – Leibler, 1951)

$$\sum_{i=1}^{k} p(\delta_i \mid Z,D) \; log \frac{p(\delta_i \mid Z, D)}{p(\delta_i \mid Z,\theta)}$$

cuyo valor medio debemos minimizar. Obviamente, esto es equivalente a maximizar

$$\int p(Z \mid D) \sum_{i=1}^{n} p(\delta_i \mid Z,D) \; log \; p(\delta_i \mid Z, \theta) \, dZ \qquad (9)$$

esto es, aproximado por Monte Carlo la integral (9), se trata de encontrar el valor θ del vector θ del vector θ que maximiza

$$\frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{k} p(\delta_i \mid Z_j \mid D) \; log \; p(\delta_i \mid Z_j, \theta)$$

en el caso de datos prospectivos, y del que maximiza

$$\sum_{i=1}^{k} p(\delta_i) \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{i=1}^{k} p(\delta_i \mid Z_j,D) \; log \; p(\delta_i \mid Z_j, \theta)$$

en el caso de datos retrospectivos.

Una vez obtenido el valor óptimo de $\hat{\theta}$ con una rutina de maximización por Newton–Raphson similar a la que utilizan los programas de regresión logística, la distribución diagnóstica aproximada correspondiente a un nuevo paciente con atributos Z viene simplemente definida por

$$l_i = log \frac{p(\delta_i \mid Z,D)}{p(\delta_k \mid Z,D)} = \hat{\theta}_{oi} + \hat{\theta}_i' Z \,, \qquad i = 1, \ldots, k-1 \qquad (10)$$

de forma que la expresión lineal $\hat{\theta}_{oi} + \hat{\theta}_i' Z$ da directamente, en la escala $(-\infty, \infty)$, una medida de la probabilidad relativa del desorden δ_i respecto del desorden δ_k tomado como referencia. Naturalmente, estas medidas pueden reconvertirse fácilmente en probabilidades mediante las expresiones

$$p(\delta_i \mid Z,D) = exp(l_i) \; p(\delta_k \mid Z,D) \,, \qquad i = 1, \ldots, k-1$$
$$p(\delta_k \mid Z,D) = 1/\{1 + \sum_{i=1}^{k-1} exp(l_i)\} \qquad (11)$$

Es interesante subrayar que la distribución diagnóstica obtenida no depende del desorden δ_k arbitrariamente elegido como referencia.

Ejemplo (cont.)

En la Tabla 3 se recogen los resultados de aplicar el modelo propuesto (8) y su aproximación (11) a los cuatro nuevos casos de síndrome de Cushing descritos en la Tabla 2, cuyo verdadero origen, posteriormente confirmado mediante histopatología, aparece en la última columna

j	$\lambda_1 Z$	λ_2	$p(\delta_i \mid Z,D)$			$p(\delta_i \mid Z, \hat{\theta})$			δ
22	10 5570	3 1732	.1299	.8683	.0018	.2805	.7186	.0010	2
23	20.4267	9 6409	.0043	.1337	.8620	.0005	.1394	.8600	3
24	18 1632	6 8603	.0113	.9550	.0337	.0036	.9527	.0437	2
25	3 9577	−.8382	.7218	.2776	.0006	.9166	.0834	.0000	1

Tabla 3: *Distribuciones diagnósticas*

Como puede observarse, tanto la distribución diagnóstica exacta como sus aproximaciones, asignan probabilidades altas al verdadero desorden. El valor óptimo de la matriz de parámetros logísticos, que resulta ser

$$\hat{\theta} = \begin{pmatrix} 14.2361 & -6.7571 & -2.6792 \\ 7.0945 & -1.7978 & -2.6820 \end{pmatrix}$$

permite obtener con facilidad la distribución diagnóstica aproximada para cualquier otro caso. Así, si un paciente presenta en la orina 9 0 y 1 50 mg/29 h de tetrahidrocortisona y pregnanetriol respectivamente, los logaritmos de los cocientes de sus probabilidades diagnósticas serán

$$l_1 = 14.2361 - 6.7571 \; log \; (9\,0) - 2.6792 \; log \; (1\,5) = -1\,6971$$

$$l_2 = 7.0945 - 1.7978 \; log \; (9.0) - 2.6820 \; log \; (1\,5) = +2\,0569$$

lo que ya indica que δ_1 y δ_2 son bastante menos y bastante más probables respectivamente que δ_3. La distribución diagnóstica aproximada resultará ser

$$p(\delta_i \mid Z, \hat{\theta}) = (0.0203 \,, 0\,8686 \,, 0.1111)$$

4. DISCUSION

El modelo descrito permite obtener soluciones razonables en un amplio espectro de situaciones, es fácilmente programable, y parece superar las limitaciones más importantes de los métodos alternativos descritos en la literatura.

De hecho, la teoría de *funciones de evaluación propias* (Savage, 1971) permite demostrar (Bernardo, 1983) que los resultados obtenidos, tanto con datos simulados como con problemas reales, son notablemente mejores que los proporcionados por el análisis discriminante y por la regresión logística.

Las limitaciones de espacio que nos hemos impuesto solo nos permite mencionar el importante problema de la selección de atributos en los problemas de diagnóstico automático.

Se trata de un problema de decisión secuencial cuya solución, bajo condiciones muy generales (Bernardo & Bermudez, 1985) consiste en seleccionar aquel conjunto de atributos que minimiza el valor medio de la *entropia* de las distribuciones diagnósticas a que da lugar,

$$\int p\,(\mathrm{Z}\mid \mathrm{D}) \sum_{i=1}^{k} p\,(\delta_i \mid \mathrm{Z,D})\, \log p\,(\delta_i \mid \mathrm{Z,D})\, d\,\mathrm{Z} \qquad (12)$$

En la práctica, se parte de un subconjunto de atributos considerados especialmente relevantes, se eliminan progresivamente aquellos cuya supresión no aumenta el valor de (12), y se procede entonces a añadir secuencialmente aquellos atributos cuya incorporación disminuye el valor de (12), hasta alcanzar el máximo que el problema permita.

REFERENCIAS

AITCHISON J y DUNSMORE I. R. (1975): *Statistical Prediction Analysis* Cambridge: University Press.

ANDERSON J. A. (1972): Separate sample logistic discrimination *Biometrika* 67, 217–272.

ANDERSON T. W (1951): Clasification by multivariate analysis. *Psychometrika* 16, 31–50.

BERMUDEZ J. D (1979): *Modelos de Decisión Predictiva con Aplicaciones Médicas* Tesis de Licenciatura Universidad de Valencia

BERMUDEZ, J. D (1984): *Modelos de Clasificación Regulares* Tesis Doctoral Universidad de Valencia.

BERNARDO J. M (1978): Métodos Bayesianos y diagnóstico clínico *Estadist. Española* 78/79, 39–56

BERNARDO J. M (1979 a): Expected information as expected utility *Ann. Statist.* 7, 686–690

BERNARDO J. M. (1979 b): Reference posterior distributions for Bayesian Inference *J. Roy Statist Soc.* B 41, 113–147 (con discusión).

BERNARDO, J. M. (1980): El concepto de aproximación en la metodología estadística. *Rev Acad. Ci Madrid* 74, 307–309

BERNARDO J. M (1983): Bayesian logistic diagnostic distributions. *Tech Rept* 8/83. Dept Bioestadística, Univ Valencia.

BERNARDO J. M. y BERMUDEZ J. D. (1985): The choice of variables in probabilistics classification. *Bayesian Statistics 2* (Bernardo, J. M., DeGroot, M. H., Lindley, D. V. y Smith, A.F.M., eds.) 67–81. Amsterdam: North Holland (con discusión).

COX D. R. (1970): *The Analysis of Binany Data* London: Methuen

DAWID A. P (1976): Properties of diagnostic data distributions. *Biometrics* 32, 647–658

DE GROOT M. H (1970): *Optimal Statistical Decisions.* New York: Mc Graw – Hill

DIACONIS, P. y FREEDMAN D (1984): Asymptotics of graphical projection punsuit. *Ann Statist.* 12, 793–815

FISHER, R. A. (1936): The use of multiple measurements in taxonomic problems. *Ann Eugen* 7, 179–188

GEISSER, S. (1964): Posterior odds for multivariate normal classification *J. Roy Statist Soc* B 26, 69–76

GOEL, P. K. (1983): Information measures and Bayesian hieraschical models *J Amer Statist Assoc* 78, 408–410.

KULLBACK S & LEIBLER R. A. (1951): On information and suffiensy *Ann. Math Statist* 22 525–540.

PRESS S. J. and WILSON S. (1978): Chosing between logistic regression and discriminant analysis. *J Amer Statist. Assoc.* 73, 699–705.

SAVAGE, L. J. (1971): Elecitation of personal probabilities and expectations *J Amer Statist Assoc* 66, 783–801.

TITTERINGTON et al (1981): Comparison of discrimination techniques applied to a complex data set of head injured patients. *J Roy Statist Soc* A 144, 145–175 (con discusión)

WALD A. (1945): On a statistical problem arising in the classification of an individual in one of two groups. *Ann Math. Statist.* 15, 145–163.

SUMMARY

AUTOMATIC DIAGNOSIS IN MEDICINE

This paper describes a procedure to derive the probabilities associated to each possible disease of a particular patient, given his or her symptoms and a data set containing the symptoms and final classification of past patients. The model acknowledges the interdependence among symptoms and lies completely within a Bayesian framework.

Key words Discriminant analysis; probabilistic classification; diagnostic distributions; reference distributions; Bayesian methods; logistic regression.

AMS 1980 Subject Classification: 62H30, 62A10.