Métodos estadísticos de clasificación

Preparado por Luis M. Molinero (Alce Ingeniería)
CorreoE: **bioestadistica@alceingenieria.net**

Artículo en formato PDF

Diciembre 2002
www.seh−lelha.org/stat1.htm

El problema de la clasificación es uno de los primeros que aparecen en la actividad científica y constituye un proceso consustancial con casi cualquier actividad humana, de tal manera que en la resolución de problemas y en la toma de decisiones la primera parte de la tarea consiste precisamente en clasificar el problema o la situación, para después aplicar la metodología correspondiente y que en buena medida dependerá de esa clasificación. Por supuesto también es así en la medicina, ciencia en la que el diagnóstico constituye una parte primordial, siendo una fase previa para la aplicación de la terapia. Diagnosticar es equivalente a *clasificar* a un sujeto en una patología concreta en base a los datos correspondientes de su anamnesis, exploración y pruebas complementarias. Cuando hablamos de clasificar a un sujeto en un grupo determinado, a partir de los valores de una serie de parámetros medidos u observados, y esa clasificación tiene un cierto grado de incertidumbre, resulta razonable pensar en la utilización de una metodología probabilística, que nos permita cuantificar esa incertidumbre.

Desde el punto de vista estadístico podemos distinguir dos enfoques diferentes al problema de la clasificación. En el primero de ellos los grupos están bien definidos y se trata de determinar un criterio para etiquetar cada individuo como perteneciente a alguno de los grupos, a partir de los valores de una serie limitada de parámetros. En este caso las técnicas más utilizadas se conocen con el nombre de **análisis discriminante**, aunque como veremos existen otras posibles alternativas, tales como la utilización de la **regresión logística**. El segundo enfoque corresponde a aquel caso en el que a priori no se conocen los grupos y lo que precisamente se desea es establecerlos a partir de los datos que poseemos. Ahora tenemos en esencia un problema taxonómico, y las técnicas estadísticas más utilizadas en ese área se conocen con el término **análisis de "cluster"**, que podemos traducir como **análisis de agrupaciones** y también como **análisis de conglomerados** por algunos autores.

Uno de los problemas más simples en cuanto a metodología de clasificación es el de etiquetar a un sujeto como enfermo SI/NO, a partir del resultado de una prueba diagnóstica, que ya fue tratado en un artículo reciente. Pero en casi cualquier actividad, no solo científica, es raro que las cosas sean tan simples y que se maneje una sola variable para tomar la decisión clasificadora; lo habitual será disponer de un conjunto de variables, y entonces resulta ideal utilizarlas de forma conjunta, lo que nos conduce a un enfoque multivariante de la cuestión.

Análisis discriminante

En el análisis discriminante estudiamos las técnicas de clasificación de sujetos en grupos ya definidos. Partimos de una muestra de *N* sujetos en los que se ha medido *p* variables cuantitativas independientes, que son las que se utilizarán para tomar la decisión en cuanto al grupo en el que se clasifica cada sujeto, mediante el modelo matemático estimado a partir de los datos. Dentro del análisis discriminante nos encontramos a su vez con dos enfoques diferentes, uno que denominaremos **predictivo** y otro **explicativo**.

En el **análisis discriminante predictivo** se trata de estimar a partir de los datos unas ecuaciones que aplicadas a un nuevo sujeto, para el que se determinan los valores de las diferentes variables, pero del que se desconoce a qué grupo pertenece, nos proporcionen una regla de clasificación lo más precisa posible. Se trata pues de formular un algoritmo por el que se pueda determinar a qué grupo pertenece una nueva observación. Este tipo

de análisis puede constituir por ejemplo una ayuda al diagnóstico, o un método de ayuda a la decisión sobre la utilización de una terapia concreta. En el análisis discriminante predictivo es importante cuantificar con qué precisión se clasificará a un nuevo sujeto.

A diferencia del anterior, en el **análisis discriminante descriptivo** estamos más interesados en las variables empleadas para diferenciar los grupos, en las variables explicativas, y lo que deseamos es determinar cuáles de esas variables son las que más diferencian a los grupos, cuales son importantes y cuales no a efectos de clasificar los sujetos.

Mediante las ecuaciones estimadas en el procedimiento de análisis discriminante obtenemos un mecanismo para asignar un sujeto a uno de los grupos, a partir de los valores de las variables explicativas. Si estamos trabajando sólo con dos grupos, en la asignación existen dos posibles errores: el que se comete al clasificarlo en el primer grupo, cuando en realidad pertenece al segundo P(I/II), y el que se cometería al incluirlo en el segundo grupo, cuando en realidad pertenece al primero P(II/I). El criterio matemático de clasificación se determina de tal manera que minimice la probabilidad de error, que en el caso más general de prevalencias diferentes en cada grupo con valores P(I) y P(II), será

$$P(error) = P(I/II)\ P(II) + P(II/I)\ P(I)$$

Cuando la importancia de cada uno de los errores es diferente, por ejemplo si estamos ante un diagnóstico, cuando es más grave el error que se comente al clasificar a un individuo enfermo como sano (falso negativo) que el que se cometería al clasificar a uno sano como enfermo (falso positivo), el criterio de clasificación puede tenerlo en cuenta, introduciendo en la ecuación que se va a minimizar un peso o coste para cada error. Si llamamos C1 al peso o coste del error de clasificar en el grupo II a un sujeto del grupo I, y C2 al de clasificar en el grupo I a un sujeto del grupo II, se trata ahora de minimizar la ecuación

$$C(error) = C1\ P(I/II)\ P(II) + C2\ P(II/I)\ P(I)$$

siendo C1+C2=1

Cuando tenemos dos grupos y *p* variables explicativas, el análisis discriminante nos permite estimar los coeficientes *b0, b1... bp* de una función de clasificación

$$D = b_0 + b_1 \cdot x_1 + b_1 \cdot x_1 + \ldots + b_p \cdot x_p$$

Siendo P(I) y P(II) las prevalencias de cada grupo, C1 y C2 los costes de clasificación incorrecta anteriormente definidos, y si denominamos

$$c = \frac{C_2 \cdot P(II)}{C_1 \cdot P(I)}$$

la regla de decisión consiste en clasificar una observación concreta de ***X*** en el grupo I cuando D > ln c, y clasificarla en el grupo II cuando D < ln c

Este procedimiento se generaliza para clasificación en más de dos grupos.

La regresión logística como herramienta de análisis discriminante

El principal inconveniente del análisis discriminante tradicional radica en que supone que los grupos pertenecen a poblaciones con distribución de probabilidad normal multivariante para las variables explicativas

X1 a *Xp*, con igual matriz de varianzas y covarianzas. Por ello no debiera incluirse en el modelo variables que no cumplieran esa condición, lo que no permite por ejemplo la utilización de variables cualitativas.

Sin embargo, en el modelo de regresión logística, que se explicaba en un artículo anterior, se estima la probabilidad de un suceso en función de un conjunto de variables explicativas y en la construcción del mismo no hay ninguna suposición en cuanto a la distribución de probabilidad de esas variables, por lo que pueden intervenir variables no normales y variables cualitativas. Si tenemos dos grupos, de tal manera que un sujeto o pertenece al grupo I o al II (por ejemplo o tiene hipertensión o no la tiene), podemos considerar el modelo de regresión logística como una fórmula para calcular la probabilidad de pertenecer a uno de esos grupos, y estimar así la probabilidad de que una observación X pertenezca al grupo I, o su complementaria la probabilidad de que pertenezca al grupo II. De esta forma podemos considerar la regresión logística como una alternativa al análisis discriminante. Además la interpretación del resultado de aplicar una ecuación logística es más intuitiva al tratarse de un valor de probabilidad.

Cuando tenemos más de dos grupos el modelo logístico introducido para una variable dicotómica (dos grupos) se puede extender de forma natural, con pocas modificaciones, conociéndose entonces como **regresión logística politómica**. La idea es la siguiente: en la regresión logística si denominamos *P* a la probabilidad del suceso (por ejemplo probabilidad de pertenecer al grupo I), su complementaria (probabilidad de no suceso o en nuestro caso probabilidad de pertenecer al grupo II) es *1−P*. Al cociente entre ambas probabilidades se le denomina odds del suceso, y el modelo logístico para dos categorías postula

$$odds = \frac{p}{1-p} = \frac{P(I/\mathbf{X})}{P(II/\mathbf{X})} = \exp(b_0 + b_1 x_1 + \ldots + b_p x_p)$$

es decir que modela el odds del grupo I respecto al grupo II. *P(I/**X**)* indica probabilidad condicionada a observar el vector ***X***.

Consideremos ahora tres grupos, es decir un modelo logístico para una variable cualitativa con tres posibles categorías, y referenciemos el primer grupo con el valor 0, el segundo con el 1 y el tercero con 2, para utilizar la terminología habitual de la regresión. Podemos extender el modelo anterior de la forma siguiente:

$$o_1(\mathbf{X}) = \frac{P(1/\mathbf{X})}{P(0/\mathbf{X})} = \exp(b_{10} + b_{11} x_1 + \ldots + b_{1p} x_p)$$

$$o_2(\mathbf{X}) = \frac{P(2/\mathbf{X})}{P(0/\mathbf{X})} = \exp(b_{20} + b_{21} x_1 + \ldots + b_{2p} x_p)$$

que modela el odds del grupo 1 respecto al 0 y del grupo 2 respecto al 0. A partir de esos odds se puede calcular la probabilidad condicional de cada una de las categorías (probabilidad de pertenecer a cada uno de los grupos), dado ***X*** unos valores de las variable explicativas. Después de unas sencillas operaciones algrebraicas obtenemos:

$$P(0/\mathbf{X}) = \frac{1}{1 + \exp(o_1) + \exp(o_2)}$$

$$P(1/\mathbf{X}) = \frac{\exp(o_1)}{1 + \exp(o_1) + \exp(o_2)}$$

$$P(2/\mathbf{X}) = \frac{\exp(o_2)}{1 + \exp(o_1) + \exp(o_2)}$$

Estimación de la tasa de aciertos de clasificación

Una medida intuitiva sobre la calidad de un método de clasificación lo constituye la tasa de aciertos, proporción de observaciones bien clasificadas, y resulta adecuado conocer cuánto mayor es la tasa de aciertos obtenida con respecto a la que es previsible esperar de una clasificación aleatoria.

El sistema ideal de validar un modelo de clasificación sería aplicarlo a otra muestra obtenida de la misma población y determinar en la nueva muestra la tasa de aciertos utilizando el algoritmo clasificador; pero es una situación que no se suele presentar en la práctica, ya que implica un diseño más costoso al precisar de dos muestras, una para estimar y otra para validar. Por otro lado, si calculamos la tasa de aciertos clasificatorios en la propia muestra que ha sido utilizada para construir el modelo, es fácil comprender que ese valor constituye una estimación optimista de la tasa de aciertos real, que es la que nos interesa estimar, sobre todo si se desea utilizar el modelo con fines predictivos. La razón de ese sesgo radica en que lógicamente se ha estimado el modelo de tal manera que se minimice la probabilidad de clasificación errónea en esas observaciones.

Un método menos optimista para validar el modelo utilizando los mismos datos, se basa en acudir a una técnica similar a la empleada en el cálculo de estimadores jacknife, consistente en trabajar con la muestra a excepción de una de las observaciones y estimar en esa nueva muestra de tamaño $N-1$ el modelo y a partir de él clasificar la unidad que no intervino en los cálculos. Repitiendo el proceso para las N observaciones, sacando una observación diferente cada vez, obtenemos una tasa de aciertos con menor sesgo.

Selección de variables a utilizar para discriminar entre los grupos

En casi cualquier análisis multivariante nos encontramos con la necesidad de seleccionar variables: identificar las variables más relacionadas con el resultado que se estudia y qué variables no parece que guarden relación. Y el análisis discriminante, o en general la construcción de modelos para clasificación, no constituyen una excepción a este deseo. Conviene, una vez más, hacer una llamada de atención en cuanto a la tendencia a utilizar **técnicas estadísticas automáticas de selección de variables**, conocidas como métodos de **regresión por pasos o "stepwise"**, que lamentablemente son utilizadas con excesiva frecuencia, quizás porque los programas modernos permiten su fácil empleo y favorecen la pereza intelectual, a pesar de la amplia crítica que suscitan entre los expertos. Hemos seleccionado dos enlaces por su interés en relación con este asunto.

Análisis de agrupaciones (cluster)

En este tipo de análisis, a diferencia del anterior, sólo disponemos de los valores de p variables $\boldsymbol{X}$ explicativas, para N sujetos, y el objetivo es agruparlos en K grupos ($K \leq N$), de tal manera que los individuos que pertenecen a un grupo se parezcan lo más posible entre sí con respecto a esas variables, y a su vez difieran lo máximo posible de los individuos de otros grupos. Este planteamiento es completamente diferente de la metodología estadística habitual ya que aquí no hay una hipótesis previa. Un posible ejemplo puede ser el buscar grupos de procesos médicos para valoración de costes, de tal manera que los grupos sean lo más homogéneos en cuanto a los recursos empleados.

Existen diferentes procedimientos para construir los grupos, y diferentes formas de determinar cómo se mide la similitud. Para ello se introduce el concepto de **distancia** entre las observaciones, que a su vez también viene determinado por el tipo de variables que se analizan, ya sean éstas cuantitativas como por ejemplo la presión arterial, cualitativas ordinales en las que al resultado se le puede asignar un número cuyo orden tiene

sentido, pero no la diferencia entre dos valores, y cualitativas nominales que corresponden a una etiqueta y donde la similitud se determina como simple coincidencia de valores.

Cuando se analizan sólo dos variables los datos son representables en unos ejes XY y de forma visual se puede intentar determinar una posible formación de grupos, por lo que una sencilla técnica a emplear es buscar, mediante algún método de reducción de variables (por ejemplo análisis de componentes principales), obtener dos nuevas variables, función de las originales, que conserven una gran parte de la variabilidad original, y representarlas gráficamente para una inspección visual.

Aunque en ocasiones encontramos análisis de agrupaciones en la literatura biomédica, no es una técnica muy habitual. Actualmente, un área en la que se está utilizando con cierta frecuencia es en epidemiología, en estudios geográficos de riesgos y de distribución de enfermedades, con el fin de determinar si existen agrupaciones sospechosas de casos, fundamentalmente buscando relación con entornos contaminados o con focos contaminantes. Remitimos al lector interesado al artículo sobre "Disease mapping" referenciado en los enlaces.

Enlaces de interés

- ***Cluster analysis and disease mapping−−why, when, and how? A step by step guide***
  Sjurdur F Olsen, Marco Martuzzi, and Paul Elliott
  BMJ 1996; 313: 863−866. [Full text]
- ***Commentary: Classification and cluster analysis***
  B S Everitt
  BMJ 1995; 311: 535−536. [Full text]
- ***Selección algorítmica de modelos en las aplicaciones biomédicas de la regresión múltiple***
  Luis Carlos Silva Ayçaguer, Isabel Barroso Utra
  Medicina Clínica 116(19): 741−745 (España); 2001
- ***Use of Stepwise Methodology in Discriminant Analysis***
  Jean S. Whitaker
  Texas A&M University, January 1997
  Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, January, 1997

Indice de artículos

Principio de la página