METODOLOGIA PROPUESTA PARA EL CALCULO DE LA FUNCION DISCRIMINANTE OPTIMA

Dunia Chávez[1], T. Burgos[2] y Nydia Hernández[3]
[1]Departamento de Matemática, Universidad Agraria de La Habana (UNAH)
[2]Departamento de Matemática Aplicada e Informática, Centro Nacional de Sanidad Agropecuaria
[3]Departamento de Estadística, Universidad de La Habana

RESUMEN
El trabajo fue realizado en el Centro Nacional de Sanidad Agropecuaria (CENSA) y el Departamento de Matemáticas de la Universidad Agraria de La Habana (UNAH), en el año 1997. Se empleó una de las técnicas estadísticas multivariadas usada frecuentemente en la práctica: el Análisis Discriminante. Se realizó un estudio de las diferentes reglas discriminantes en dos poblaciones con variables aleatorias continuas, presentes en varios sistemas estadísticos, para establecer una metodología de cálculo de la función lineal discriminante óptima. Se incluye además la estimación del error de clasificación por los métodos de Reclasificación y Validación Cruzada. Finalmente dos ejemplos con datos reales obtenidos en investigaciones del CENSA completan el estudio desarrollado.

Palabras clave: análisis discriminante, sistemas estadísticos, metodología, función óptima, error de clasificación.

MSC: 62P10.

ABSTRACT
This paper was developed in the Centro de Sanidad Animal (CENSA) and the Departament of Matematics of Universidad Nacional Agrícola (UNAH) in 1997). One of the most frequently used multivariate statistical technique was employed: Discriminant Analysis. The existing different discriminant rules were studied in two populations with continous random variables, which are included in different statistical packages, for establishing a methodology for calculating the optimum linear discriminant function. The classification error for reclassification and crossvalidation methods are included. The analysis of the developed study is completed with two examples with real data.

1. INTRODUCCION

En la actualidad, el análisis estadístico es un complemento básico de casi todas las investigaciones del presente. Según Anderson (1958), en múltiples situaciones de la vida no puede hacerse uso de los métodos matemáticos determinísticos para conocer la verdad, ya que el conocimiento del fenómeno es incompleto y están presentes factores aleatorios que no permiten explicarlo. De ahí que en el desarrollo de un gran número de ciencias particulares, los métodos estadísticos jueguen un papel fundamental.

Cooley and Lohnes (1971) señalan que en varios procesos se involucran un gran número de variables con comportamientos disímiles, de aquí el surgimiento del análisis multivariado como el método esencial para resolver problemas en que estén involucrados varios indicadores.

Debido a algunas investigaciones biológicas y antropométricas surgen las primeras ideas del análisis discriminante en la cuarta década del siglo XX. Esta es una de las técnicas multivariadas de mayor aplicación dentro de la estadística y se encarga de clasificar un individuo en una de las poblaciones dadas a priori, a partir de la información brindada por muestras aleatorias de cada una de ellas (Hernández (1997)).

Es conocido que los paquetes de computadoras han estado disponibles durante largo tiempo para ayudar en el análisis de datos y que en muchos de ellos se encuentran procedimientos para el análisis discriminante (Nachtshein (1997)). De esta manera se obtiene un conjunto de funciones discriminantes de las cuales resultaría interesante seleccionar la óptima. Motivado por esta problemática el Departamento de Matemática Aplicada e Informática del CENSA y el Departamento de Matemática de la UNAH se proponen realizar este trabajo con el objetivo de establecer una metodología para el cálculo de la función lineal discriminante óptima. En ella se utilizan las diferentes funciones discriminantes calculadas por los sistemas estadísticos actuales y su correspondiente estimación del error de clasificación por los métodos de Resustitución y/o Validación Cruzada. Esta metodología es comparada con el análisis clásico y se discuten los resultados obtenidos al aplicar la misma, a dos ejemplos prácticos con datos reales del CENSA.

2. METODOLOGIA DE TRABAJO PARA EL CALCULO DE LA FUNCION DISCRIMINANTE OPTIMA

Debido a que el Departamento de Bacteriología del Centro Nacional de Sanidad Agropecuaria utiliza con mucha frecuencia la técnica de Análisis Discriminante y con el objetivo de sintetizar la búsqueda de una función discriminante para un problema determinado en los diferentes paquetes estadísticos, se estableció la siguiente metodología de trabajo a fin de obtener una función discriminante que fuera óptima y a su vez lo más rápido posible

Paso 1: Organizar los datos y calcular funciones discriminantes utilizando diferentes sistemas estadísticos por el método de selección paso a paso. Este método se propone porque en ocasiones el investigador no desea que todas las variables estén presentes en el modelo sino que estén las que más contribuyan a la explicación del problema en cuestión.

Paso 2: Calcular la estimación del error de clasificación para cada una de las funciones discriminantes por los métodos de Resustitución y/o Validación Cruzada. Estos se obtienen en el Paso 1.

Paso 3: Seleccionar la función discriminante óptima teniendo en cuenta que el error de clasificación estimado en el Paso 2 sea mínimo.

3. COMPARACION DE LA NUEVA METODOLOGIA PROPUESTA CON EL ANALISIS CLASICO. RESULTADOS OBTENIDOS A PARTIR DE BASES DE DATOS REALES

En la literatura se reportan varios tipos de funciones discriminantes como son: la de Fisher, la cuadrática, la logística, la canónica, la factorial y la que se calcula a través de la regresión pero en la práctica, el investigador pretende encontrar una de ellas que resuelva su problema sin tener en cuenta de qué tipo pueda ser. Esto lo realiza utilizando cualquier paquete estadístico, perdiéndose así la posibilidad de que existan otras funciones discriminantes "mejores", y que a su vez expliquen también el problema. Como el parámetro que mide la bondad de dichas funciones es el error de clasificación se estableció en el Paso 3 seleccionar la función discriminante cuyo error sea lo más pequeño posible. Por ello consideramos que a través de la nueva metodología se obtiene una eficiente función discriminante. A continuación se presentan dos problemas prácticos con datos reales obtenidos por investigadores del CENSA:

1. Se tiene un grupo de 115 ovinos a los cuales se le inocularon cepas de *Corynebacterium pseudotuberculosis;* se le tomaron mediciones de anticuerpos en sangre o las llamadas células sanguíneas. Este grupo fue analizado por una técnica denominada ELISA, la cual brinda la posibilidad de saber si el animal se enfermó o no (Hunt, Kezar and Honman (1996)). Estas mediciones son continuas y fueron las siguientes:

H_1 = HEMOGLO	(hemoglobina)
H_2 = HEMATOC	(hematocritos)
H_3 = LEUTOT	(leucocitos totales)
X_4 = NEUTROF	(neutrófilos)
X_5 = LINFOC	(linfocitos)

La variable X_6 se denomina ELISA. Es una variable binaria que toma el valor 1 si el ovino se enfermó y 0 en caso contrario.

2. Se tiene un grupo de 84 cerdos a los cuales se le midieron algunos indicadores corporales y otros como son variables climáticas y condiciones del lugar donde se encontraban. Estas variables fueron:

X_1 = PESO	(peso)
X_2 = ESTATURA	(estatura)
X_3 = TEMP	(temperatura ambiental)
X_4 = HR	(humedad relativa)
X_5 = HUMD	(humedad del local).

La variable X_6 se denomina DIARR. Es una variable binaria que toma el valor 1 si el cerdo tuvo diarrea y 0 en caso contrario.

Luego en ambos casos se está en presencia de un problema de dos poblaciones con variables aleatorias continuas. EL objetivo de los investigadores que plantearon el problema era encontrar una función que les permitiera decidir a qué clase pertenece un ovino o un cerdo del cual poseen sus datos reales. Como se puede apreciar, ambos son casos donde inicialmente se puede buscar una posible solución a través del análisis discriminante. A continuación se ofrecen los resultados de cada uno de los problemas planteados en dos tablas diferentes. Estos datos fueron obtenidos a través de los diferentes sistemas estadísticos. Además se brinda una solución final para cada uno, teniendo en cuenta que el error de clasificación sea mínimo.

Tabla 1. Resultados del análisis discriminante para el ejemplo 1./
Discriminant analysis results for the example 1.

Sist. Estad.	Tipo de Función	Función Discriminante	Error Resust.	Error Cross. Val.
CSS [4]	canónica ascend.	$F(X) = 08194\ X_1 + 0.1478\ X_3 + 2.9903\ X_4 - 13.4964$	0.3228	-
SAS [13]	canónica cuadrat.	$F(X) = 0.2251\ X_1^2 - 1355.6887\ X_2^2 - 0.0151\ X_3^2 + 65.7661\ X_4^2 + 40.5044\ X_5^2 + 27.4616\ X_1X_2 - 0.0724\ X_1X_3 + 24.1864\ X_1X_4 + 25.5003\ X_1X_5 + 10.8444\ X_2X_3 - 824.6925\ X_2X_4 - 764.6828\ X_2X_5 + 4.4804\ X_3X_4 + 3.7982\ X_3X_5 + 105.5024\ X_4X_5 - 37.2350\ X_1 + 1236.2820\ X_2 - 6.3892\ X_3 - 167.4065\ X_4 - 175.4617\ X_5 - 315.1997$	0.3053	-
STAT GRAPHIS	canónica	$F(X) = 1.0565\ X_1 - 9.6015\ X_2 + 0.1524\ X_3 + 6.5491X_4 + 4.1558\ X_5 - 16.6575$	0.3492	-
SPSS	canónica paso a paso, asc, desc.	$F(X) = 0.8558\ X_1 - 10.5630$	0.3478	-
STATIS TICA	Fisher ascend.	$F(X) = 0.7032\ X_1 + 0.1268\ X_3 + 2.5672\ X_4 - 11.4840$	0.3228	-
SYSTAT	regresión	$F(X) = - 0.046\ X_1 + 0.049\ X_3 + 0.964\ X_5 + 0.028$	0.5	-
STATITCF	factorial	$F(X) = 1.2013\ X_1 - 0.2896\ X_2 + 0.3329\ X_3 + 0.6770\ X_4 + 0.4124\ X_5$	0.3492	-
AD [14]	selecc. de vars.	$F(X) = 0.9242\ X_1 + 0.1672\ X_3 + 2.7264\ X_5 - 8.0981\ X_2 + 12.02$	0.2879	0.2083

Tabla 2. Resultados del análisis discriminante para el ejemplo 2./
Discriminant analysis results for the example 2.

Sist. Estad.	Tipo de Función	Función Discriminante	Error Resust.	Error Cross. Val.
CSS [4]	canónica ascend.	$F(X) = 0.7283\ X_1 + 0.1522\ X_3 + 2.9001\ X_5 - 10.2553$	0.3131	-
SAS [13]	canónica cuadrat.	$F(X) = 0.2134\ X_1^2 - 1222.657\ X_2^2 - 0.1512\ X_3^2 + 50.131\ X_4^2 + 38.0324\ X_5^2 + 25.3216\ X_1X_2 - 0.0524\ X_1X_3 + 21.1254\ X_1X_4 + 20.5311\ X_1X_5 + 10.2004\ X_2X_3 - 800.6115\ X_2X_4 - 664.6265\ X_2X_5 + 3.1045\ X_3X_4 + 2.7235\ X_3X_5 + 101.2421\ X_4X_5 - 32.2221\ X_1 + 1330.1829\ X_2 - 5.9298\ X_3 - 159.4511\ X_4 - 165.4141\ X_5 - 301.9713$	0.2987	-
STAT GRAPHIS	canónica	$F(X) = 0.7023\ X_1 - 7.6125\ X_2 + 0.1324\ X_3 + 5.9110\ X_4 + 2.8998\ X_5 - 11.1227$	0.3221	-
SPSS	canónica paso a paso	$F(X) = 0.6988\ X_1 - 10.1630$	0.3312	-
STATIS TICA	Fisher ascend.	$F(X) = 0.7011\ X_1 + 0.1223\ X_3 + 2.8672\ X_5 - 11.2110$	0.3116	-
SYSTAT	regresión	$F(X) = 0.124\ X_1 + 0.129\ X_3 + 1.325\ X_5 + 1.198$	0.4980	-
STATITC	factorial	$F(X) = 0.9024\ X_1 - 0.2655\ X_2 + 0.2088\ X_3 + 0.6990\ X_4 + 2.870\ X_5$	0.3302	-
AD [14]	selecc. de vars.	$F(X) = 0.8777\ X_1 + 0.1335\ X_3 + 2.7190\ X_4 - 7.0765\ X_2 - 9.7874$	0.2695	0.2101

Como se puede apreciar, las soluciones óptimas serían en ambos casos, las funciones calculadas por el sistema AD por poseer menor error de clasificación.

Este sistema AD propuesto por Suárez (1996), es la ampliación de un sistema de cómputo de reglas discriminantes que estaba realizado para dos poblaciones con variables aleatorias continuas, discretas y mezcla de ambas, a un sistema de Análisis Discriminante para k poblaciones y los mismos tipos de variables. El sistema se realizó sobre Windows 95 y permite la clasificación de individuos a través de la función discriminante de Fisher, la cuadrática y la logística. Esto es una diferencia de los paquetes comerciales conocidos en nuestro país donde en la mayoría se calcula la función discriminante de tipo canónica. Además AD incluye el cálculo de la estimación del error de clasificación por ambos métodos mencionados anteriormente; mientras que solamente en el paquete estadístico SAS (1982) lo hace igual, los demás solo por Resustitución.

Un aspecto muy particular que destaca este sistema entre otros es que en el cálculo de todas las funciones utiliza un procedimiento de selección de variables que toma como criterio, minimizar la estimación

del error de clasificación. En los otros sistemas la selección de variables la realizan con el uso de pruebas de hipótesis asumiendo normalidad, sin embargo, esta suposición en la práctica no siempre se cumple.

Cuando iniciamos este trabajo y establecimos la metodología se llegó a la función óptima seleccionándola de las soluciones parciales de cada uno de los sistemas con la condición antes mencionada.

No obstante, se destaca que el sistema AD aunque no sea un paquete estadístico para todo tipo de análisis de datos, si es eficiente para el AD por lo que se recomienza su uso en el caso que necesite aplicar dicha técnica multivariada.

CONCLUSIONES

1. Se estableció una metodología de trabajo para el cálculo de la función discriminante lineal óptima.

2. Como complemento de esta primera conclusión, se hace un estudio de los diferentes sistemas estadísticos respecto al Análisis Discriminante. De esta manera se ofrece a los investigadores de todas las ramas la posibilidad de dar soluciones más completas a las aplicaciones prácticas de esta teoría. No obstante, se debe señalar que en la mayoría de los sistemas estudiados se encuentran las técnicas de Análisis Discriminante más antiguas como son la idea del factorial y la canónica, sin embargo, no aparecen valores importantes de esta temática como son: la estimación del error, la selección de variables, etc.; los cuales constituyen parámetros para medir la bondad del discriminante. En los sistemas más actuales (A..D. Suárez, (1996)) donde aparece la selección de variables sí se controló la estimación del error de clasificación, no así en los anteriores este procedimiento se realiza a través de una prueba F en la cual se asume normalidad.

3. En ambos problemas planteados se logra encontrar una solución final en cada uno de los sistemas, de las cuales el investigador puede escoger la que desee según su interés. Además se obtuvieron las funciones óptimas:

$$F_1(X) = 0.9242\ X_1 + 0.1672\ X_3 + 2.7264\ X_4 - 8.0981\ X_2 + 12.02$$

$$F_2(X) = 0.8777\ X_1 + 0.1335\ X_3 + 2.7190\ X_5 - 7.0765\ X_2 - 9.7874$$

REFERENCIAS

[1] ANDERSON, T.W. (1958): "Introduction to multivariate statistical analysis", John Wiley, New York, 121-142.

[2] BELLO, A.L. (1995): "On the performance of rank transforms discriminant method in error-rate estimation", **J. Statist. Comput. Simul**. 48:153-165.

[3] COOLEY, W.W. and P.R. LOHNES (1971): **Multivariate Data Analysis**, John Wiley, New York, 299-323.

[4] CSS: STATISTICA. StatSoft, 1991.

[5] CUADRAS, C.M. (1991): "Métodos de Análisis Multivariante. Estadística y Análisis de Datos", Promociones y Publicaciones Universitarias S.A., Barcelona.

[6] HERNANDEZ, N.P. (1997): "Regla de clasificación para dos poblaciones normales con matriz de covarianza muestral singular", **Revista Investigación Operacional** 13(1):19-25.

[7] HULLINGER, G.A.; D.P: KNOWLES; T.C: McGUIRE and W.P: CHEEVERS (1995): "Ovine Arthritis-Encephalitis Lentivirus SU is the Ligand for infection of ovine Synovial Membrane Cells", **Virology** 192(2):328-332.

[8] KHIRSAGAR, A.M. (1972): "Multivariate analysis", Marcel Dekker, New York, 187-244.

[9] HUNT, C.W.; W. KEZAR and D.D: HONMAN (1996): "Effects of Hybrid and Ensiling with a Microbial Inoculant on the Inmunological characteristics of ovine animals", **Journal of Animal Science** 71(1):38-44.

[10] KNOKE, D.J. (1996): "The robust estgimation of classification. Error rates", **Comp. & Maths. Appls.** 12 A(2):253-260.

[11] LINARES, G. (1994): **Análisis de datos**, MES, Universidad de La Habana, Facultad Matemática-Cibernética, Cuba. 321-345.

[12] NACHTSHEIN, C.J. (1997): "Tools for Computer-Aided Design of Experiments", **Journal of Quality Technology** 19(3):132-160.

[13] SAS, SAS Institute Inc. **Introduction to SAS**, User´s Guide: Statistics (1982).

[14] SUAREZ, I. (1996): "Sistema de Cómputo de Análisis Discriminante", Trabajo de Diploma, Universidad de La Habana.

[15] TARDIF, B. and J. HARDY (1996): "Assessing the relative contribution of variables in canonical discriminant analysis", **Biometrics** 44(1):69-76.

[16] VARELA, M. (1996): "Aplicación de la estadística multivariada en las ciencias agropecuarias", Programa y Resúmenes X Seminario Científico INCA. Cultivos Tropicales 17(2):21-26.