Revista Colombiana de Estadística
Diciembre 2009, volumen 32, no. 2, pp. 247 a 265

Comparación entre tres técnicas de clasificación

Comparison for three Classification Techniques

Freddy Hernández Barajas[1,a], Juan Carlos Correa Morales[2,b]

[1]Departamento de estadística, Instituto de Matemática y Estadística, Universidad de São Paulo, São Paulo, Brasil

[2]Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Medellín, Colombia

Resumen

En este artículo se muestran los resultados de un estudio de comparación mediante simulación de tres técnicas de clasificación, regresión logística multinomial (MLR), análisis discriminante no métrico (NDA) y análisis discriminante lineal (LDA). El desempeño de las técnicas se midió usando la tasa de clasificación errónea. Se encontró que las técnicas MLR y LDA tuvieron un desempeño similar y muy superior a NDA cuando la distribución multivariada de las poblaciones es normal o logit-normal; en el caso de distribuciones multivariadas log-normal y $Sinh^{-1}$-normal la técnica MLR tuvo mejor desempeño.

***Palabras clave*:** regresión logística multinomial, análisis discriminante no métrico, análisis discriminante lineal, clasificación.

Abstract

In this paper we show the results of a comparison simulation study for three classification techniques: Multinomial Logistic Regression (MLR), No Metric Discriminant Analysis (NDA) and Linear Discriminant Analysis (LDA). The measure used to compare the performance of the three techniques was the Error Classification Rate (ECR). We found that MLR and LDA techniques have similar performance and that they are better than DNA when the population multivariate distribution is Normal or Logit-Normal. For the case of log-normal and $Sinh^{-1}$-normal multivariate distributions we found that MLR had the better performance.

***Key words*:** Logistic regression, Nonparametric discriminant analysis, Multiple classification.

[a]Estudiante de doctorado. E-mail: fhernanb@ime.usp.br
[b]Profesor asociado. E-mail: jccorrea@unalmed.edu.co

1. Introducción

El proceso de asignar una observación p variada en uno de varios grupos preestablecidos se denomina clasificación. El objetivo básico es construir una función discriminante que tome la información de las p variables para resumirla en un indicador con el cual se pueda clasificar la observación de manera correcta en uno de los grupos. En la literatura estadística se pueden encontrar varios métodos desarrollados para abordar el problema de clasificación.

Una de las técnicas más conocidas en clasificación fue propuesta por Fisher (1936); este enfoque se denomina Análisis Discriminante Lineal (LDA) y básicamente divide el espacio muestral en subespacios mediante hiperplanos que permiten separar lo mejor posible los grupos en estudio. Los supuestos para la utilización de LDA son: normalidad multivariada e igualdad de matrices de covarianzas entre los grupos. Welch (1939) mostró la optimalidad de LDA bajo condiciones de normalidad multivariada. Rao (1948) propuso el análisis discriminante canónico el cual es una generalización del análisis discriminante lineal para el caso de varios grupos. Clunies & Riffenburgh (1960) y Anderson (1972) encontraron una función discriminante para el caso donde no se cumple el supuesto de igualdad de matrices de covarianzas.

En estimación recientemente Hawkins & McLachan (1997) y Croux & Dehon (2001) propusieron un procedimiento llamado high breakdown para remover observaciones atípicas que pueden afectar las estimaciones de los vectores de medias y de las matrices de covarianzas para cada uno de los grupos utilizados por LDA y que, por tanto, pueden afectar el proceso de clasificación. Cheng et al. (2002) propusieron dos formas alternativas de realizar las estimaciones del vector de medias y de la matriz de covarianzas cuando hay datos faltantes; las propuestas consisten básicamente en combinar el algoritmo ER (Estimation-Robust) propuesto por Little & Smith (1987) con los estimadores high breakdown.

Otra de las técnicas de clasificación se debe a Raveh (1983 y 1989) quien propuso el Análisis Discriminante no Métrico (NDA), el cual utiliza como punto de partida la función discriminante generada por LDA y por medio de un proceso iterativo propuesto por Choulakian & Almhana (2001) se mejora la función discriminante inicial; la calidad de la nueva función discriminante se mede por medio de un índice de separación entre los grupos; este índice fue propuesto por Guttman (1998). La ventaja que tiene NDA sobre LDA es que NDA no requiere supuestos distribucionales.

Otra de las técnicas de clasificación es el modelo de regresión logística, sugerido por Cornfield (1962), Cox (1966) y Day & Kerridge (1967) para una variable con respuesta binaria; Anderson (1972) propuso el modelo de regresión logística multinomial o policotomo (MLR). Pregibon (1981) planteó un método para detectar posibles observaciones atípicas en la muestra de entrenamiento que se usa para encontrar los estimadores de máxima verosimilitud para un modelo de regresión logístico. Trevor & Ferry (1991) presentaron un nuevo modelo de regresión logística robusto que mostró tener mejor desempeño en un estudio de simulación y en un estudio con datos reales. Carroll & Pederson (1993) mostraron que existen otras

dos formas de estimaciones resistentes y robustas que tienen sesgos iguales o menores en las estimaciones cuando se tienen observaciones atípicas en el conjunto de observaciones. Una ventaja de la herramienta MLR es que no requiere supuestos distribucionales.

1.1. Técnicas de clasificación

Supóngase que se tiene la tarea de clasificar observaciones (sujetos u objetos) p variadas en uno de g grupos ya establecidos y que para esto se cuenta con un conjunto de observaciones proveniente de cada uno de los grupos. Sean $x_{ij} \in \Re^p$ las observaciones donde $j = 1, \ldots, g$, $i = 1, \ldots, n_j$ tal que n_j representa el número de observaciones que pertenecen al grupo j. Este conjunto de n observaciones[1] se denomina conjunto de entrenamiento y permite la construcción de la función discriminante para cada una de las técnicas. A continuación se describe de cada una de las tres técnicas estudiadas.

1.1.1. Análisis discriminante lineal

Sean $\overline{x}_j$ y S_j los vectores de medias y las matrices de covarianzas de cada uno de los g grupos y sea $\overline{x}$ el vector de medias global del conjunto de entrenamiento. Usando estos elementos se pueden definir dos matrices B y W que representan la variabilidad entre los grupos y la variabilidad dentro de los grupos respectivamente y que están dadas por:

$$B = \sum_{j=1}^{g} n_j (\overline{x}_j - \overline{x})(\overline{x}_j - \overline{x})' \tag{1}$$

$$W = \sum_{j=1}^{g} (n_j - 1) S_j \tag{2}$$

El objetivo de la técnica LDA es encontrar un vector $a \in \Re^p$ de tal manera que se maximice el cociente Λ definido por (3); así se encuentra un hiperplano que genera la máxima diferencia entre la variabilidad intergrupal e intragrupal.

$$\Lambda = \frac{a'Ba}{a'Wa} \tag{3}$$

Rencher (1998) muestra que los valores de a que maximizan Λ se pueden estimar por medio de los vectores propios $e_1, e_2, \ldots, e_s$ asociados a los valores propios positivos[2] $\lambda_1 > \lambda_2 > \cdots > \lambda_s$ de $W^{-1}B$. De esta manera si $\widehat{a} = e_1$ entonces $\hat{a}$ se denomina primer discriminante lineal estimado (LD_1) y este corresponde a la función discriminante. La regla para clasificar una nueva observación x en uno de

[1] En total $n = \sum_{j=1}^{g} n_j$.

[2] Donde $s = min\{p, g-1\}$.

los grupos basada en el primer discriminante lineal consiste en asignar x al grupo j si se cumple que $\left[\widehat{a}'(x-\overline{x}_j)\right]^2$ es mínimo.

1.1.2. Análisis discriminante no métrico

La técnica NDA propuesta por Raveh (1989) se basa en los resultados de LDA y consiste en la búsqueda de $\eta \in \Re^p$ que maximice a *Disco* (coeficiente de discriminación) propuesto por Guttman (1998) dado en (4).

$$Disco = \frac{\sum_{k=1}^{g}\sum_{h=1}^{g} n_k n_h \left|\eta'\left[\overline{x}_k - \overline{x}_h\right]\right|}{\sum_{k=1}^{g}\sum_{h=1}^{g}\sum_{i=1}^{n_k}\sum_{l=1}^{n_h}\left|\eta'\left[x_{ik} - x_{lh}\right]\right|} \tag{4}$$

El numerador de (4) representa la separación entre los diferentes grupos mientras que el denominador representa la variación total entre todas las observaciones. Una propiedad importante es que $0 \leq Disco \leq 1$, donde el valor de cero se obtiene solamente cuando todos los vectores de medias para cada grupo del conjunto de entrenamiento son iguales y el valor de uno cuando los puntajes[3] para cada grupo al ser puestos sobre una recta quedan completamente separados unos de otros, es decir, no se observan traslapes entre los puntajes de grupos diferentes.

Con base en los g vectores de medias y los n vectores de observaciones se pueden definir dos tipos de matrices de orden $p \times p$ así:

$$B_{kh} = \left[\overline{x}_k - \overline{x}_h\right]\left[\overline{x}_k - \overline{x}_h\right]' \tag{5}$$

$$V_{kh}(i,l) = \left[x_{ik} - x_{lh}\right]\left[x_{ik} - x_{lh}\right]' \tag{6}$$

De esta manera *Disco* en (4) puede representar en forma matricial como una función del vector η:

$$Disco(\eta) = \frac{\eta' B(\eta)\eta}{\eta' V(\eta)\eta} \tag{7}$$

donde $B(\eta)$ y $V(\eta)$ son matrices simétricas de orden $p \times p$ que dependen del parámetro η de la siguiente manera:

$$B(\eta) = \sum_{k=1}^{g}\sum_{h=1}^{g} \frac{n_k n_h B_{kh}}{\sqrt{\eta' B_{kh}\eta}} \tag{8}$$

$$V(\eta) = \sum_{k=1}^{g}\sum_{h=1}^{g}\sum_{i=1}^{n_k}\sum_{l=1}^{n_h} \frac{V_{kh}(i,l)}{\sqrt{\eta' V_{kh}(i,l)\eta}} \tag{9}$$

Para maximizar *Disco* en (4) con respecto a η, Choulakian & Almhana (2001) propusieron el siguiente algoritmo:

[3]El puntaje z de una observación x por medio de η se define como $z = \eta' x$.

1. Comenzar con $\eta_0 = \theta^*$, siendo θ^* el vector propio[4] (LD_1).
2. Calcular $\eta_{k+1} = \eta_k\,[1 - 2Disco(\eta_k)] + 2V(\eta_k)^{-1}B(\eta_k)\eta_k$, para $k = 0, 1, 2, \ldots$
3. Detener el proceso cuando $|Disco(\eta_{k+1}) - Disco(\eta_k)| \leq \epsilon$, donde ϵ es un valor real positivo definido con anterioridad; por ejemplo, $\epsilon = 10^{-5}$.
4. Obtener el valor óptimo de la función discriminante η haciendo $\eta = \eta_k$.

Luego de encontrar el valor óptimo de η, este puede utilizarse como función discriminante para clasificar nuevas observaciones en uno de los g grupos. Para realizar la clasificación se determinan $g-1$ puntos de corte (CP) de la siguiente manera: se toman los n puntajes z_{ij} del conjunto de entrenamiento con $i = 1, 2, \ldots, n_j$ y con $j = 1, 2, \ldots, g$; sin pérdida de generalidad, se puede suponer que los primeros n_1 puntajes son menores que los segundos n_2 y así sucesivamente. El punto de corte CP_1 que separa los grupos 1 y 2 es igual al percentil $100(n_1/n)\,\%$ de los n puntajes ordenados, el segundo CP_2 que separa los grupos 2 y 3 es igual al percentil $100((n_1 + n_2)/n)\,\%$ de los n puntajes ordenados; de manera similar se obtienen los $G - 3$ CP restantes. La regla para clasificar una nueva observación x en uno de los grupos es calcular el puntajes z de x por medio de η y clasificar la nueva observación x al grupo 1 si $z \leq CP_1$, al grupo k si $CP_{k-1} < z \leq CP_k$ o al grupo g si $z > CP_{g-1}$.

1.1.3. Regresión logística multinomial

La técnica MLR es un caso particular de modelos lineales generalizados donde la variable respuesta y corresponde a una variable aleatoria independiente multinomial cuya media se modela por medio de un conjunto de covariables. Para el problema de clasificación el valor de la variable respuesta corresponde a $y = 1, \ldots, g$ y el conjunto de covariables se denota por $\mathrm{x}' = (1, x_1, \ldots, x_p)'$ que conforman el conjunto de entrenamiento.

Sea π_j la probabilidad de que una observación x pertenezca al grupo $j = 1, \ldots, g$, uno de los grupos es considerado como grupo de referencia, sin pérdida de generalidad, se puede tomar el grupo 1 como referencia; entonces el logit para cada uno de los otros grupos se define como:

$$logit(\pi_j) = log\left(\frac{\pi_j}{\pi_1}\right) = x'\beta_j \quad , \quad j = 2, \ldots, g \tag{10}$$

Una vez estimados los $g - 1$ vectores β_j usando el conjunto de entrenamiento, es posible estimar la probabilidad de que una nueva observación x pertenezca a uno de los grupos usando las expresiones 11 y 12 y asignar la observación al grupo que tenga mayor probabilidad.

$$\widehat{\pi}_1 = \frac{1}{1 + \sum_{j=2}^{g} exp\big(x'\widehat{\beta}_j\big)} \tag{11}$$

[4]Obtenido mediante Análisis Discriminante Lineal.

$$\widehat{\pi}_j = \frac{exp\big(x'\widehat{\beta}_j\big)}{1 + \sum_{j=2}^{g} exp\big(x'\widehat{\beta}_j\big)} \quad , \quad j = 2, \ldots, g \tag{12}$$

1.2. Estudios previos de simulación

En la literatura estadística se pueden encontrar estudios que comparan el desempeño de las técnicas análisis discriminante lineal, regresión logística y análisis discriminante no métrico. A continuación se muestran los principales resultados obtenidos en estos estudios.

1.2.1. Comparaciones entre LDA y regresión logística (LR)

Efron (1975) comparó las dos técnicas para el caso de dos grupos con igual matriz de covarianzas y encontró que la eficiencia relativa asintótica (ARE) de LR con respecto a LDA está entre un medio y dos tercios. Crawley (1979) comparó LR con LDA para muestras pequeñas con dos grupos y encontró que para el caso de matrices de covarianza iguales LDA tiene un mejor desempeño que LR en el proceso de clasificación, para el caso de matrices de covarianzas diferentes LR tuvo ligeramente un mejor desempeño y para el caso de dos poblaciones distribuidas no normal LR tuvo un desempeño muy superior a LDA. Harrell & Lee (1985) realizaron una comparación entre las técnicas para el caso de dos grupos considerando normalidad con matrices de covarianzas iguales, tamaños de muestra de 50 y 130 con seis distancias de Mahalanobis entre los vectores de medias de las dos poblaciones que variaron entre los valores de 0.94 y 4.68; en este estudio se encontró que el desempeño de LDA fue mejor que LR pero que las diferencias no eran significativas.

Pohar et al. (2004) llevaron a cabo un estudio donde compararon LDA y LR por medio de simulación. Para comparar los desempeños de cada una de las técnicas utilizaron el índice típico de tasa de clasificación errónea y los índices A, B, C y Q propuestos por Harrell & Lee (1985). La comparación se inició en un escenario en el cual se cumplían los supuestos de LDA y luego realizaron cambios en los tamaños de muestra, matriz de covarianzas y distancia de Mahalanobis entre las medias de los grupos simulados. Se encontró que los desempeños de LDA y LR fueron muy cercanos, siempre y cuando los supuestos de normalidad no sean afectados fuertemente, y presentaron lineamientos para identificar este tipo de situaciones; adicionalmente, discutieron las situaciones donde es inapropiado utilizar LDA para clasificación.

1.2.2. Comparaciones entre LDA y MLR

Shelley & Donner (1987) llevaron a cabo un estudio de comparación con el objetivo de extender los resultados de Efron (1975) para medir la eficiencia relativa asintótica (ARE) de regresión logística multinomial con respecto a LDA para

el caso de poblaciones distribuidas normal multivariada con igual matriz de covarianzas. Los casos que estudiaron consideraron dos, tres y cuatro grupos. Los autores encontraron que en el caso de vectores de pendientes logístico colineales ARE estaba en el intervalo de 50 % a 65 % para dos grupos y de 35 % a 95 % para el caso de cuatro grupos cuando la distancia de Mahalanobis entre el grupo de referencia y los demás estuvo en 3.0 a 3.5. Para el caso de vectores ortogonales se encontró que ARE decae rápidamente a medida que aparecen más grupos en el proceso de clasificación.

1.2.3. Comparaciones entre LDA y NDA

Raveh (1989) llevó a cabo un estudio de simulación donde comparó LDA y NDA para el caso de dos grupos; se consideraron tres escenarios o tipos de distribuciones de probabilidad multivariada para cada grupo: normal multivariada, log-normal y chi-cuadrada; en cada uno de estos escenarios se consideraron distribuciones bivariadas y trivariadas. El tamaño de muestra fue siempre de 50 observaciones para cada uno de los dos grupos. El objetivo básico del estudio fue comparar el desempeño de las dos técnicas usando la tasa de clasificación errónea para el conjunto de entrenamiento y para un nuevo conjunto de validación obtenidos de la misma distribución.

Para el caso de dos grupos provenientes de una distribución normal multivariada (2 ó 3 variables) se encontró que cuando hay igualdad entre las matrices de covarianzas LDA tiene tasas de clasificación erróneas como máximo 1 % mejores que las de NDA. Se encontró también que a medida que las matrices de covarianza difieren entre sí, la ventaja de LDA disminuye hasta el punto que NDA obtiene menor tasa de clasificación errónea para el caso extremo de matrices de covarianza. Para el caso de grupos provenientes de una distribución log-normal se encontró que NDA es muy superior que LDA; el desempeño de NDA estuvo por encima de LDA en 16 % para conjuntos de entrenamiento y 14 % para conjuntos de validación. Para este mismo caso se halló que, a medida que los parámetros de la distribución log-normal para cada grupo difieren, el desempeño de NDA mejora sobre el de LDA. Para el caso de grupos provenientes de una distribución chi-cuadrado se encontró que NDA tuvo un desempeño similar a NDA; las diferencias entre las tasas de clasificación erróneas fueron 1 % a favor de NDA; se observó también que la ventaja de NDA sobre LDA se incrementaba ligeramente a medida que disminuían los grados de libertad de la distribución.

Choulakian & Almhana (2001) realizaron una comparación entre LDA y NDA usando tres conjuntos de datos: poultry data, encontrado en Raveh (1983), conformado por diez grupos con cuatro variables; wolf skull data, encontrado en Morrison (1990), conformado por cuatro grupos y nueve variables, y feelings data, encontrado en Hand (1989), conformado por cuatro grupos y veinticinco variables. En cada una de estas tres aplicaciones se construyó la función discriminante NDA con base en la función discriminante de LDA y se encontró que NDA clasifica mejor el conjunto de entrenamiento, también se halló un aumento en el coeficiente de discriminación Disco para cada una de las aplicaciones: para la primera aplicación

Disco pasó de 0.9915 a 0.9935, para la segunda *Disco* cambió de 0.987 a 1 y para la última *Disco* pasó de 0.9615 a 0.9832.

2. Objetivo del estudio

La tarea de clasificar sujetos u objetos en grupos preestablecidos dado un conjunto de características siempre ha sido un problema a resolver en diferentes ámbitos; algunos ejemplos son: asignación de créditos, determinación del estado de enfermedad de un paciente, envío de publicidad para clientes de una empresa, entre otros. Las características o variables observadas en los sujetos a clasificar pueden ser cuantitativas y/o cualitativas, y en caso de que existan variables cualitativas las condiciones de normalidad multivariada no están aseguradas; por tanto, es importante estudiar y comparar técnicas que puedan incorporar la información de variables cualitativas, lo cual es el caso de Análisis Discriminante no Métrico y Regresión Logística. Por otra parte, como se observa en la sección anterior, no existen en la literatura estadística estudios de comparación entre LDA, NDA y MLR para el caso de más de dos grupos.

El objetivo principal de este artículo es estudiar mediante simulación el desempeño de las tres técnicas LDA, NDA y MLR en el proceso de clasificación para el caso de más de dos grupos bajo diferentes distribuciones de probabilidad multivariada.

Se decidió incluir LDA en el estudio para extender los trabajos de Raveh (1989) a más de dos grupos y los ejemplos de Choulakian & Almhana (2001) ya que la función discriminante NDA se basa en el primer discriminante de LDA.

2.1. Metodología

La comparación de las técnicas NDA y MLR se llevó a cabo por medio de simulaciones en las cuales se consideraron tres tipos de distribuciones de probabilidad multivariada para los grupos, diferentes parámetros y varios tamaños de muestra.

2.1.1. Escenarios

Los escenarios considerados en el estudio fueron cuatro:

Escenario 1. Tres grupos normales bivariados (denotados por 1, 2 y 3), matrices de covarianzas iguales donde $\sigma_1 = \sigma_2 = 1$ con valores de correlación ρ de 0.1, 0.3, 0.5, 0.7 y 0.9. Se consideraron cinco situaciones de alejamiento gradual para los grupos; el grupo 1 siempre fue el de referencia y estuvo ubicado en el origen del plano cartesiano mientras que las ubicaciones de los otros dos grupos cambiaron sobre los ejes. A representa la situación más cercana de vectores de medias y E la más lejana.

- Situación A: $\mu_1 = (0, 0)$, $\mu_2 = (1, 0)$, $\mu_3 = (0, 1)$.
- Situación B: $\mu_1 = (0, 0)$, $\mu_2 = (1, 0)$, $\mu_3 = (0, 2)$.

- Situación C: $\mu_1 = (0,0)$, $\mu_2 = (1,0)$, $\mu_3 = (0,3)$.
- Situación D: $\mu_1 = (0,0)$, $\mu_2 = (2,0)$, $\mu_3 = (0,2)$.
- Situación E: $\mu_1 = (0,0)$, $\mu_2 = (2,0)$, $\mu_3 = (0,3)$.

Los tamaños muestrales utilizados fueron de 20, 50, 100 y una combinación de estos últimos.

Escenario 2. Este escenario es similar al anterior con respecto a las situaciones de media para cada grupo y tamaños muestrales; la diferencia está en que se consideraron matrices de covarianzas diferentes para cada grupo.

La matriz de covarianzas del grupo 1 se caracterizó por $\sigma_1 = \sigma_2 = 1$ con diferentes valores de correlación ρ de 0.1, 0.3, 0.5, 0.7 y 0.9. La matriz de covarianzas de los grupos 1, 2 y 3 se denotan por Σ_1, Σ_2 y Σ_3. Los casos considerados fueron:

- $\Sigma_2 = 2\Sigma_1$ y $\Sigma_3 = 2\Sigma_1$.
- $\Sigma_2 = 2\Sigma_1$ y $\Sigma_3 = 3\Sigma_1$.
- $\Sigma_2 = 3\Sigma_1$ y $\Sigma_3 = 3\Sigma_1$.
- $\Sigma_2 = 2\Sigma_1$ y $\Sigma_3 = 4\Sigma_1$.

Escenario 3. Siete grupos distribuidos normal trivariada. Se consideraron tres situaciones de alejamiento gradual para los grupos; el grupo 1 siempre fue el de referencia y estuvo ubicado en el origen mientras que las ubicaciones de los otros grupos cambiaron de manera gradual. A representa la situación más cercana y C la más lejana.

- Situación A: $\mu_1 = (0,0,0)$, $\mu_2 = (0,0,2)$, $\mu_3 = (0,0,-2)$, $\mu_4 = (0,2,0)$, $\mu_5 = (0,-2,0)$, $\mu_6 = (2,0,0)$, $\mu_7 = (-2,0,0)$.
- Situación B: $\mu_1 = (0,0,0)$, $\mu_2 = (0,0,2)$, $\mu_3 = (0,0,-2)$, $\mu_4 = (0,3,0)$, $\mu_5 = (0,-3,0)$, $\mu_6 = (3,0,0)$, $\mu_7 = (-3,0,0)$.
- Situación C: $\mu_1 = (0,0,0)$, $\mu_2 = (0,0,2)$, $\mu_3 = (0,0,-2)$, $\mu_4 = (0,4,0)$, $\mu_5 = (0,-4,0)$, $\mu_6 = (4,0,0)$, $\mu_7 = (-4,0,0)$.

El tamaño de las muestras tomadas de cada población fue de 10, 20 y 50. Las matrices de covarianzas de los grupos $1, 2, \ldots, 7$ se denotan por Σ_1, Σ_2,..., Σ_7. Se consideraron matrices de covarianza diferentes y la matriz de covarianzas del grupo 1 (de referencia) se caracterizó porque las varianzas de las tres variables fueron $\sigma_1 = \sigma_2 = \sigma_3 = 1$. Se consideraron valores de correlación entre pares de variables ρ de 0.1, 0.3, 0.5, 0.7 y 0.9. Las estructuras generales de las matrices de covarianzas fueron:

- Estructura 1: $\Sigma_1 = \Sigma_2 = \Sigma_3$, $\Sigma_4 = \Sigma_5 = 2\Sigma_1$, $\Sigma_6 = \Sigma_7 = 4\Sigma_1$.
- Estructura 2: $\Sigma_1 = \Sigma_2 = \Sigma_3$, $\Sigma_4 = \Sigma_5 = 4\Sigma_1$, $\Sigma_6 = \Sigma_7 = 8\Sigma_1$.

Escenario 4. Tres grupos con funciones de densidad especiales bivariadas. En este escenario se llevó a cabo la comparación de las técnicas usando tres funciones de densidad especiales: Lognormal, $Sinh^{-1}$-normal y Logit-normal. Para crear las muestras se generaron muestras de observaciones normales

bivariadas con los parámetros del escenario 1 y luego se aplicó el sistema de transformación sugerido por Johnson (1987) para obtener observaciones provenientes de cada distribución.

2.1.2. Evaluación de las funciones de clasificación

La evaluación de los desempeños de las técnicas de clasificación se llevó a cabo utilizando la Tasa de Clasificación Errónea (TCE) que se define de la siguiente manera:

$$TCE = \frac{NCE}{NOBS}$$

donde NCE corresponde al número de clasificaciones erradas por la técnica en el conjunto de validación y $NOBS$ corresponde al número de observaciones en el conjunto de validación.

2.1.3. Procedimiento de comparación

Las tres técnicas de clasificación se compararon llevando a cabo el siguiente conjunto de pasos:

1. Se simuló una muestra por cada una de las g poblaciones con las cuales se forma el conjunto de entrenamiento.
2. Con el conjunto de entrenamiento se construyeron las funciones de discriminación para LDA, NDA y MLR.
3. Se generaron nuevas muestras como en el paso 1 y con ellas se formó un nuevo conjunto llamado de validación. Todas las observaciones de este conjunto se clasificaron mediante las funciones obtenidas en el paso 2.
4. Se calculó la tasa de clasificación errónea para el conjunto de validación del paso anterior para cada técnica.
5. Los pasos 3 y 4 se repitieron mil veces y se calculó la tasa promedio de clasificación errónea para cada uno de los procedimientos de clasificación.

Las simulaciones se programaron y generon en R (R Development Core Team 2008), se usó la función `lda()` para análisis discriminante lineal y `multinom()` de la librería nnet para regresión logística multinomial; la función para análisis discriminante no métrico se programó usando el procedimiento sugerido por Choulakian & Almhana (2001) y estará disponible a petición del lector.

2.2. Resultados

Escenario 1: tres grupos normales bivariados con matrices de covarianza iguales. En este escenario se encontró que las TCE para LDA y MLR difieren en máximo 1 %, lo cual se puede observar en las figuras 1 y 2 por medio de las líneas a

trazos que se superponen. La línea continua corresponde a la TCE para NDA y siempre quedó ubicada en las figuras 1 y 2 por encima de las líneas asociadas a las otras dos técnicas; sin embargo, para valores de $\rho \geq 0.7$ las diferencias en TCE para las tres técnicas fueron de máximo 2 %. Otro patrón esperado que se encontró en este escenario fue que la TCE disminuyó a medida que se presentaron las tres situaciones siguientes: aumento de la distancia entre los vectores de medias de los grupos (situación A a E), aumento del coeficiente de correlación y aumento en el tamaño de muestra.

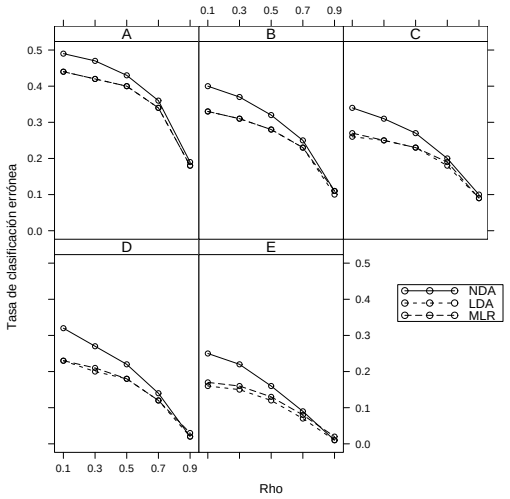

FIGURA 1: Escenario 1. Tamaños muestrales $n_1 = n_2 = n_3 = 20$.

Escenario 2: tres grupos normales bivariados con matrices de covarianza diferentes. En las figuras 3 y 4 se observan dos casos particulares del escenario y se pueden apreciar nuevamente resultados similares y los mismos patrones mencionados en el escenario anterior. Al considerar diferentes matrices de covarianzas para los grupos, los desempeños de las técnicas se vieron afectados, el aumento promedio en la tasa de clasificación errónea fue de 6 %. Se observó también que cuando las matrices de covarianzas eran diferentes con tamaños muestrales de 50 las tasas de clasificación erróneas eran prácticamente las mismas que cuando se tiene matrices de covarianzas iguales, es decir, los desempeños de las técnicas fueron similares cuando los tamaños muestrales fueron de 50 sin importar si se cumplía el supuesto de igualdad de matriz de covarianzas. En promedio LDA y MLR tuvieron tasas de clasificación errónea 3.5 % menos que NDA.

Escenario 3: siete grupos normal trivariado con matrices de covarianzas diferentes. En el presente escenario se consideraron dos estructuras de matrices de covarianzas y al pasar de la primera estructura a la segunda se afectó la TCE aumentándola. Se observó nuevamente que al aumentar el valor de ρ disminuyeron las tasas de clasificación errónea; adicionalmente, a medida que se alejan las poblaciones, la TCE disminuyó en promedio 12 %. En este escenario se comprobó

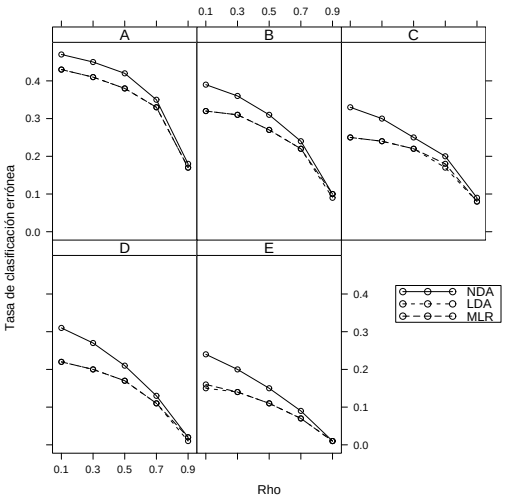

FIGURA 2: Escenario 1. Tamaños muestrales $n_1 = n_2 = n_3 = 100$.

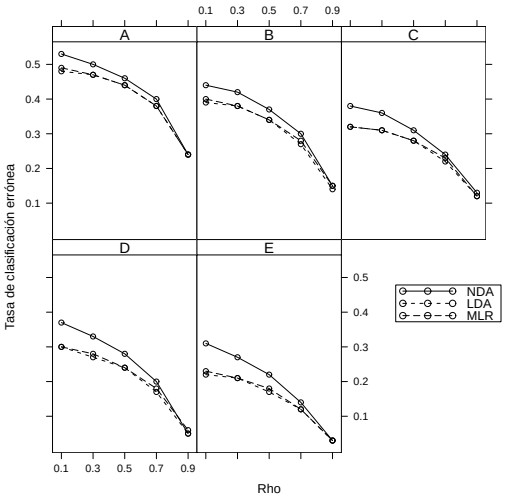

FIGURA 3: Escenario 2. Tamaños muestrales $n_1 = n_2 = n_3 = 20$ y $\Sigma_2 = \Sigma_3 = 2\Sigma_1$.

otra vez que al aumentar los tamaños muestrales se favorece el desempeño de las técnicas; en particular se encontró que TCE disminuyó en promedio 3 %. En las figuras 5, 6 y 7 se observa que el desempeño de LDA y MLR es similar y que clasifican mejor que NDA en el presente escenario.

Escenario 4: tres grupos con funciones de densidad especiales bivariadas. En la figura 8 se pueden observar las TCE para las técnicas cuando se usó la distribución lognormal con tamaños de muestra de 20. El patrón observado en las líneas de esta

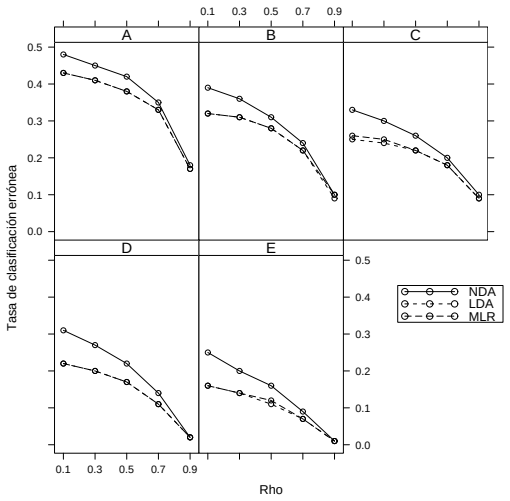

FIGURA 4: Escenario 2. Tamaños muestrales $n_1 = n_2 = n_3 = 50$, $\Sigma_2 = 2\Sigma_1$ y $\Sigma_3 = 4\Sigma_1$.

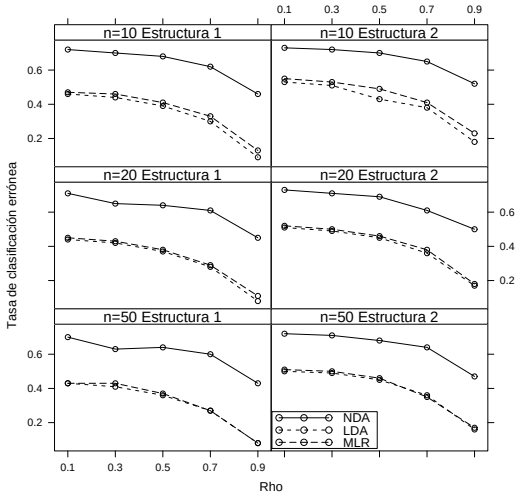

FIGURA 5: Situación A: $\mu_1 = (0,0,0)$ $\mu_2 = (0,0,2)$ $\mu_3 = (0,0,-2)$ $\mu_4 = (0,2,0)$ $\mu_5 = (0,-2,0)$ $\mu_6 = (2,0,0)$ $\mu_7 = (-2,0,0)$.

figura es igual al encontrado cuando se usaron tamaños de muestra de 50 y 100; por tanto, para este caso el aumento en el tamaño de muestra no mejoró el desempeño de las técnicas. Se encontró que LDA es la técnica con el desempeño más bajo y por lo menos el 30 % de las veces clasificó incorrectamente. Los mejores desempeños los obtuvo MLR, y a medida que aumentaba la separación entre los grupos, la diferencia con LDA y NDA fue mayor. A partir de la situación de alejamiento C la

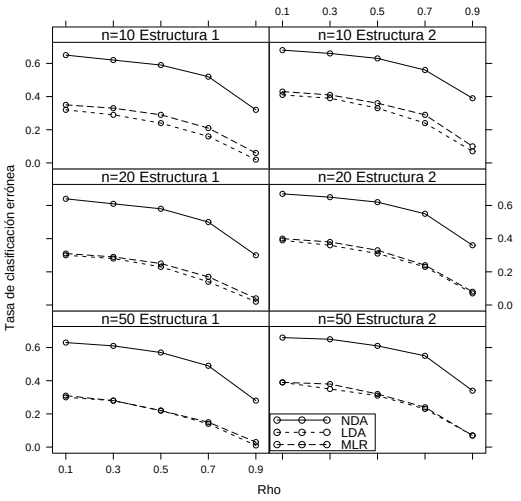

FIGURA 6: Situación B: $\mu_1 = (0,0,0)$ $\mu_2 = (0,0,2)$ $\mu_3 = (0,0,-2)$ $\mu_4 = (0,3,0)$ $\mu_5 = (0,-3,0)$ $\mu_6 = (3,0,0)$ $\mu_7 = (-3,0,0)$.

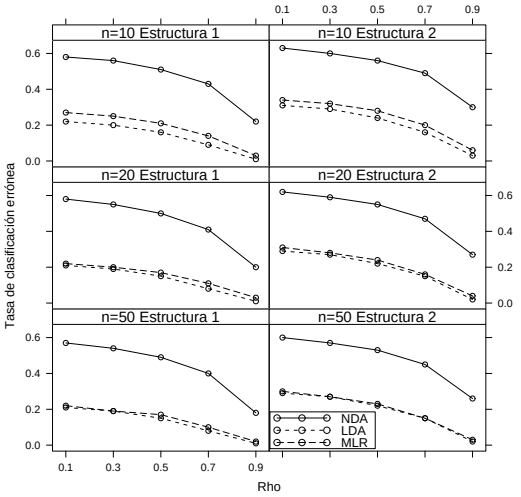

FIGURA 7: Situación C: $\mu_1 = (0,0,0)$ $\mu_2 = (0,0,2)$ $\mu_3 = (0,0,-2)$ $\mu_4 = (0,4,0)$ $\mu_5 = (0,-4,0)$ $\mu_6 = (4,0,0)$ $\mu_7 = (-4,0,0)$.

técnica LDA mantuvo su tasa de clasificación errónea independiente del coeficiente de correlación ρ mientras que para NDA y LDA se observó que el aumento en ρ mejoró el desempeño.

En la figura 9 se pueden observar las TCE para las técnicas en el caso de distribución $Sinh^{-1}$-normal. Nuevamente en este caso la técnica LDA presentó el

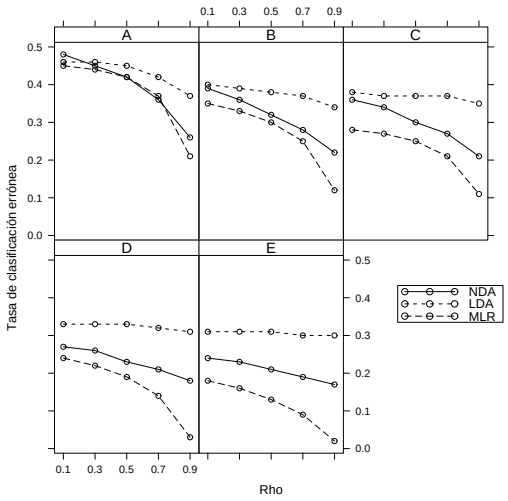

FIGURA 8: Distribución lognormal con $n_1 = n_2 = n_3 = 20$.

desempeño más bajo de las tres técnicas seguido por NDA mientras que la mejor técnica fue MLR. Los patrones mencionados anteriormente para cada una de las técnicas en el caso de la distribución lognormal se repiten nuevamente aquí; el desempeño de LDA no mejora cuando se aumenta ρ a partir de la situación D, mientras que para NDA y MLR el impacto es favorable al aumentar ρ. Se observa también que al aumentar la distancia entre los vectores de medias de los grupos las técnicas MLR y NDA mejoran el desempeño.

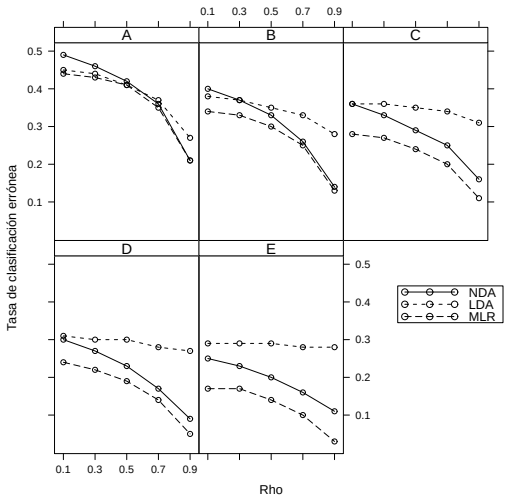

FIGURA 9: Distribución $Sinh^{-1}$-normal con $n_1 = n_2 = n_3 = 20$.

En la figura 10 se encuentran los resultados para el caso de la distribución logit-normal para tamaño muestral de 20. El patrón de las líneas de desempeño TCE para las figuras asociadas a tamaños muestrales de 50 y 100 es el mismo que el de figura 10; se encontró que al aumentar el tamaño de muestra de 20 a 50, en promedio las TCE disminuyeron 3 %, mientras que al aumentar el tamaño de muestra de 50 a 100, la disminución en TCE fue insignificante. Se observó también que la técnica NDA fue la de peor desempeño mientras que LDA y MLR tuvieron desempeños similares, la máxima diferencia de TCE entre ellas fue de 2.5 %; este último patrón fue común denominador para los escenarios 1, 2, 3 y distribución logit-normal.

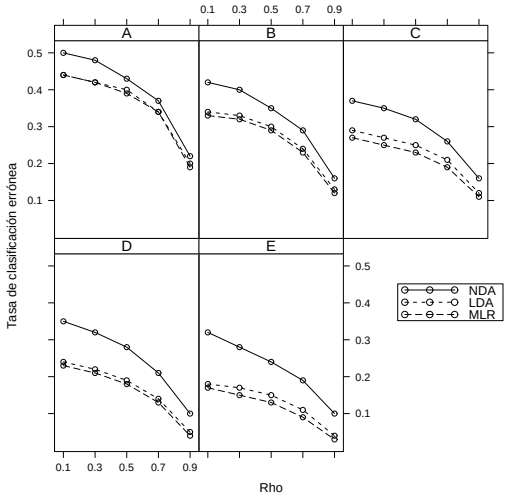

FIGURA 10: Distribución Logit-normal con $n_1 = n_2 = n_3 = 20$.

2.3. Discusión

Las técnicas de clasificación LDA y RLM tuvieron mejor desempeño al momento de clasificar en tres grupos bivariados y siete grupos trivariados; además, su desempeño fue tan similar que en la mayoría de las figuras las líneas de TCE asociadas se superponen, lo cual muestra que las diferencias son mínimas. En el estudio la técnica NDA nunca obtuvo TCE menores que MLR; los únicos casos en que NDA fue mejor que LDA fue cuando se consideraron poblaciones distribuidas lgnormal y $Sinh^{-1}$-normal.

3. Conclusiones

Hasta el momento en la literatura estadística se han reportado solamente comparaciones vía simulación entre pares de las técnicas análisis discriminante lineal,

regresión logística multinomial y análisis discriminante no métrico. En el presente estudio de simulación se compararon las tres técnicas y se encontró que análisis discriminante lineal y regresión logística multinomial tuvieron tasas de clasificación errónea muy similares y más bajas que análisis discriminante no métrico en la mayoría de los escenarios estudiados; esta última técnica presentó un mejor desempeño solo cuando la distribución poblacional fue lognormal y $Sinh^{-1}$-normal.

En situaciones prácticas donde se presente un problema de clasificación de nuevas observaciones a grupos ya definidos, teniendo en cuenta varias variables explicativas, se recomienda utilizar principalmente la técnica RLM seguida de la LDA, siempre y cuando la distribución de probabilidad de los datos sea cercana a una situación de normalidad multivariada; el supuesto de homogeneidad de varianzas puede violarse ligeramente y los resultados obtenidos con cualquiera de las técnicas serán similares; adicionalmente se sugiere utilizar este criterio siempre y cuando todas las variables explicativas sean de tipo cuantitativo.

Posibles trabajos futuros podrían encaminarse a comparar el desempeño de las técnicas considerando otro tipo de escenarios en los cuales se pueden estudiar aspectos como: mayor número de grupos a clasificar, tamaños muestrales mayores, estructuras de matrices de covarianzas diferentes, otros tipos de distribuciones para los grupos, medidas de desempeño diferentes a la tasa de clasificación errónea y algoritmos de búsqueda para determinar el vector de clasificación en la técnica NDA.

[Recibido: julio de 2008 — Aceptado: noviembre de 2009]

Referencias

Anderson, J. (1972), 'Separate Sample Logistic Discrimination', *Biometrica* **23**, 19–35.

Carroll, R. & Pederson, S. (1993), 'On Robustness in the Logistic Regression Model', *Journal of the Royal Statistical Society* **55**, 693–706.

Cheng, T., Pia, M. & Feser, V. (2002), 'High-Breakdown Estimation of Multivariate Mean and Covariance with Missing Observations', *British Journal of Mathematical and Statistical Psychology* **55**, 317–335.

Choulakian, V. & Almhana, J. (2001), 'An Algorithm for Nonmetric Discriminant Analysis', *Computational Statistics & Data Analysis* **35**, 253–264.

Clunies, C. & Riffenburgh, R. (1960), 'Geometry and Linear Discrimination', *Biometrics* **47**, 185–189.

Cornfield, J. (1962), 'Joint Dependence of the Risk of Coronary Heart Disease on Serum Cholesterol and Systolic Blood Pressure: A Discriminant Function Analysis', *Proceedings of the Federal American Society of Experimental Biology* **21**, 58–61.

Cox, D. (1966), *Some Procedures Associated with the Logistic Qualitative Response Curve*, John Wiley & Sons, New York, United States.

Crawley, D. (1979), 'Logistic Discrimination as an Alternative to Fisher's Linear Function', *New Zealand Statistician* **14**, 21–25.

Croux, C. & Dehon, C. (2001), 'Robust Linear Discriminant Analysis Using S-Estimators', *Canadian Journal of Statistics/Revue Canadienne de Statistique* **29**, 473–493.

Day, N. & Kerridge, D. (1967), 'A General Maximum Likelihood Discriminant', *Biometrics* **23**, 313–323.

Efron, B. (1975), 'The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis', *Journal of American Statistical Association* **70**, 892–898.

Fisher, R. A. (1936), 'The Use of Multiple Measurements in Taxonomic Problems', *Annual Eugenics* **7**, 179–188.

Guttman, L. (1998), 'Eta, disco, odisco and F.', *Psychometrika* **53**, 393–405.

Hand, D. (1989), *Discriminant Analysis for Psychiatric Screening*, 2 edn, John Wiley & Sons, New York, United States.

Harrell, F. E. & Lee, K. L. (1985), A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality, *in* P. K. Sen, ed., 'Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences', North-Holland, New York, United States, pp. 333–343.

Hawkins, D. & McLachan, J. (1997), 'High-Breakdown Linear Discriminant Analysis', *Journal of American Statistical Asociation* **92**, 136–146.

Johnson, M. (1987), *Multivariate Statistical Simulation*, John Wiley & Sons, New York, United States.

Little, R. & Smith, P. (1987), 'Editing and Imputing for Quantitative Survey Data', *Journal of the American Statistical Association* **82**, 58–68.

Morrison, D. (1990), *Multivariate Statistical Methods*, 3 edn, McGraw-Hill, New York, United States.

Pohar, M., Blas, M. & Turk, S. (2004), 'Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study', *Metodolski Zvezki* **1**, 143–161.

Pregibon, D. (1981), 'Logistic Regression Diagnostics', *The Annals of Statistics* **9**, 705–724.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org

Rao, C. (1948), 'The Utilization of Multiple Measurements in Problems of Biological Classification', *Journal of the Royal Statistical Society: Series B* **10**, 159–193.

Raveh, A. (1983), 'Preference Structure Analysis: A Nonmetric Approach', *Patter Recognition* **16**, 253–259.

Raveh, A. (1989), 'A Nonmetric Approach to Linear Discriminant Analysis', *Journal of the American Statistical Association* **84**, 176–183.

Rencher, A. (1998), *Multivariate Statistical Inference and Applications*, John Wiley & Sons, New York, United States.

Shelley, B. & Donner, A. (1987), 'The Efficiency of Multinomial Logistic Regression Compared with Multiple Group Discriminant Analysis', *Journal of American Statistical Association* **82**, 1118–1122.

Trevor, F. & Ferry, G. (1991), 'Robust Logistic Discrimination', *Biometrika* **78**, 841–849.

Welch, B. (1939), 'Note on Discriminant Functions', *Biometrika* **31**, 218–220.