UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2014, Vol. 7, No. 2, pp. 115–138

Alternativas de clasificación en poblaciones multivariadas

Classification alternatives in multivariate populations

Catalina Inés Cortés Vélez[a]
cicortesv@unal.edu.co

Juan Carlos Salazar Uribe[b]
jcsalaza@unal.edu.co

Resumen

Dada la importancia del tema de clasificación y los estudios que consigo se han desarrollado, en este artículo se compara, vía simulación, la eficiencia de los clasificadores Máquinas de Soporte Vectorial (SVM), Clasificador Fuzzy (FC), Regresión Logística (LR) y Análisis Discriminante Lineal (LDA), en datos provenientes de las distribuciones Normal Multivariada (MND), Skew Normal Multivariada (MSND) y t Multivariada (MTD), para diferentes números de variables. El mejor clasificador se selecciona de acuerdo con su eficiencia en términos de la tasa de clasificación errónea (TCE).

***Palabras clave*:** clasificación, distribuciones multivariadas, estadística, tasa de clasificación errónea.

Abstract

Given the importance, in the last years, of the classification topic and the study yourself have been developed, in this article we compare the efficiency of the classifiers Support Vector Machines (SVM), Fuzzy Classifier (FC), Logistic Regression (LR) and Lineal Discriminate Analysis (LDA), using Multivariate Normal Distribution (MND), Multivariate Skew Normal Distribution (MSND) and Multivariate t Distribution (MTD) for different variables number by means of a simulation study. The best classifier is selected based on your efficiency in terms of the False Discovery Rate (TCE).

***Keywords*:** clasification, false discovery rate, multivariate distribution, statistics.

[a]Estudiante de Maestría, Escuela de Estadística,Universidad Nacional de Colombia, Medellín, Colombia.

[b]Profesor asociado, Escuela de Estadística,Universidad Nacional de Colombia, Medellín, Colombia.

1. Introducción

El problema de clasificación es importante en áreas médicas, biológicas, financieras, entre otras. No hay un único método para clasificar y existen diversas propuestas reportadas en la literatura. La pregunta que surge es ¿cuál de ellos es más conveniente para clasificación?

El presente artículo tiene como tema central estudiar la eficiencia de algunos clasificadores estadísticos que permitan el apropiado agrupamiento de datos multivariados, en dos grupos, donde cada grupo posee sus propias características o parámetros que permiten, a través de los clasificadores, a cuál de los grupos pertenece cada dato.

Salazar et al. (2012) trabajan la clasificación de datos en dos grupos provenientes de poblaciones univariadas; así clasifican univariadamente usando Máquinas de Soporte Vectorial (SVM) (Cortes & Vapnik 1995), Regresión Logística (RL) (Hosmer & Lemeshow 2000) y Análisis Discriminante lineal (ADL) (Fisher 1936), cuyos clasificadores necesitan un entrenamiento previo para hacer la adecuada clasificación. En consecuencia se plantea la cuestión de analizar los clasificadores a partir de poblaciones multivariadas.

Se pretende, además de analizar estos clasificadores en poblaciones multivariadas, compararlos con el Clasificador Fuzzy o Clasificador Difuso (FC) (Hoppner et al. 1999), cuya mayor fortaleza es la de no necesitar de un grupo de entrenamiento para realizar dicha clasificación.

Se usa el método de simulación en diferentes escenarios, considerando el caso cuando se tienen dos, tres y cuatro variables en los datos provenientes de algunas distribuciones multivariadas como normal multivariada (Multivariate Normal Distribution - MND) (Anderson et al. 1958), normal asimétrica (Multivariate Skew Normal Distribution - MSND), (Azzalini & Dalla Valle 1996),(Azzalini & Capitanio 1999) y t multivariada (*Multivariate T Distribution - MTD*) (Kotz & Nadarajah 2004), presentando desde un traslapamiento fuerte hasta un traslapamiento débil. Dichas distribuciones tienen la característica en común de poseer contornos elípticos.

Para determinar la eficiencia de los clasificadores, dicha eficiencia será determinada por la tasa de clasificación errónea (TCE), que corresponde al porcentaje de datos mal clasificados.

1.1. Técnicas de clasificación

Las técnicas estadísticas de clasificación son muy utilizadas en la asignación de pertenencias a clases, cuando se cuenta con un conjunto de datos representados por $X = \{x_1, x_2, x_3, \ldots, x_n\}$, los cuales pertenecen a diferentes categorías $C_1, C_2, \ldots, C_p$, pero que de antemano no se conoce a cual categoría o clase pertenece.

En algunos casos, para los clasificadores, es necesario conocer con anticipación un

subgrupo de datos plenamente identificados en la clase que pertenece, a fin de realizar el entrenamiento de estos y proporcionar la adecuada clasificación de los datos nuevos. A continuación se da una descripción de las técnicas usadas en el presente estudio.

1.1.1. Máquinas de soporte vectorial

Se considera un conjunto de datos $\{(\boldsymbol{x}_i, y_i)\}$ para $i = 1, 2, \ldots, k$ donde $\boldsymbol{x}_i$ es el vector de datos ($\boldsymbol{x}_i \in R^n$) y y_i la clase a la que pertenece. En este estudio se tiene que $y_i \in \{-1, 1\}$ donde -1 representa la clase C_1 y 1 representa la clase C_2. El conjunto de datos en mención es denominado conjunto de entrenamiento.

En la mayoría de los casos, cuando se tienen los datos a clasificar, se encuentra que estos dos grupos están traslapados, es decir, que algunos datos de la clase C_1 pueden ser confundidos con datos de la clase C_2, lo cual hace que la margen de separación (conocida como hiperplano de separación, ver Figura 1) no sea lineal.

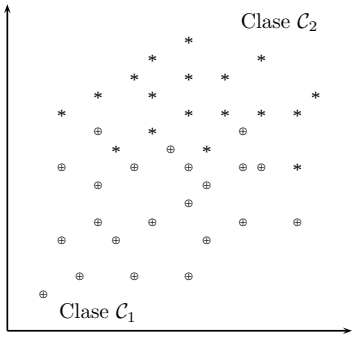

Figura 1: *dos grupos de datos, diferentes, que se traslapan. Fuente: elaboración propia.*

Este hiperplano de separación es determinado por una transformación biyectiva, no lineal $\varphi(\cdot)$, la cual hace que dentro de un espacio de dimensión mayor se convierta en un caso separable linealmente y pueda procederse de una manera menos compleja con la separación de los datos. Dicho hiperplano, en el nuevo espacio, de dimensión mayor, está dado por

$$\begin{cases} \boldsymbol{w}\varphi(\boldsymbol{x}_k) + b \geq 1 & \text{si} \quad y_k = +1 \\ \boldsymbol{w}\varphi(\boldsymbol{x}_k) + b \leq -1 & \text{si} \quad y_k = -1 \end{cases}$$

Que gráficamente corresponde a lo ilustrado en la figura 2. Así, al transformar los datos, aquellos que sean menores a -1 pertenecerán a la clase C_1, aquellos que sean menores a 1 pertenecerán a la clase C_2 y aquellos datos que sean iguales a -1 o 1 serán llamados vector soporte quienes determinan el hiperplano de separación.

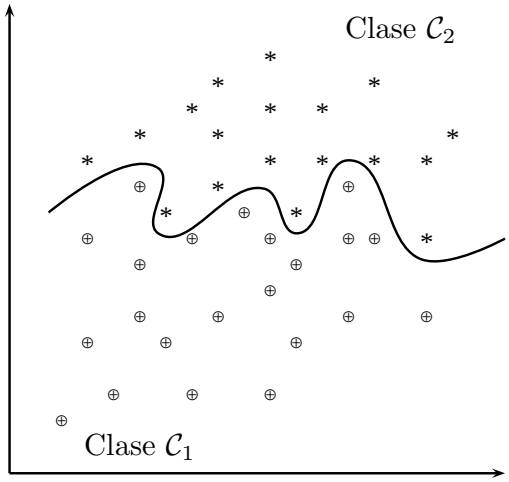

Figura 2: *hiperplano de separación entre las dos clases no separables linealmente. Fuente: elaboración propia.*

1.1.2. Clasificador difuso

Sea $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_n\}$ el conjunto de n datos a clasificar donde $\boldsymbol{x}_k$, con $k = 1, 2, \ldots, n$, es a un vector de valores correspondiente a la medición de cada variable considerada.

El número de clases en las que se desee clasificar deben ser inferiores al número de datos a clasificar, esto es, $c \leq n$ donde c es el número de clases (*clusters*) y n el número de datos.

En el clasificador difuso se tiene que cada dato a clasificar pertenece, ponderadamente, a cada clase C_i con $i = 1, 2, \cdots, c$. Así, a cada dato es asignado un vector de pertenecia a la clase, $\boldsymbol{\mu}_i((x)_k)$, y cada componente, μ_{ik}, es un valor en el intervalo $[0, 1]$. Dichas componentes deben cumplir las siguientes restricciones

$$\begin{aligned} \sum_{i=1}^{c} \mu_{ik} &= 1 \qquad & k = 1, 2, \cdots, n \\ 0 < \sum_{k=1}^{n} \mu_{ik} &< n \qquad & i = 1, 2, \cdots, c \end{aligned}$$

donde i es el índice de la clase y k es el índice para la observación.

Así, el espacio de partición difusa, M_P, estará dado por el conjunto de matrices

$$M_P = \left\{ U \;\middle/\; \mu_{ik} \in [0,1], \sum_{i=1}^{c} \mu_{ik} = 1,\ 0 < \sum_{i=1}^{n} \mu_{ik} < n \right\}$$

donde U es la matriz tal que las filas representan las clases, C_i con $i = 1, 2, \cdots, c$, y las columnas cada uno de los datos, $\boldsymbol{x}_k$ para $k = 1, 2, \cdots, n$, a clasificar. Así, un dato pertenece a la clase C_i si el valor correspondiente en la asignación matricial μ_{ik} es el mayor de todos los que están en esa columna.

Cada clase está determinada por un centro ν_{ij}, calculado a partir de la función

$$\nu_{ij} = \frac{\sum_{k=1}^{n} \mu_{ik}^{m'} x_{kj}}{\sum_{k=1}^{n} \mu_{ik}^{m'}}$$

donde m' es el parámetro que controla la cantidad de difusión en el proceso de clasificación, el cual puede tener valores en el intervalo $[1, +\infty]$, usualmente se emplea 1 o 2.

Este centro es usado para optimizar la función cuadrática media

$$J_m(U, v) = \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik}^{m'} (d_{ik})^2$$

A partir de lo anterior, la partición óptima se determina por medio del proceso iterado, fijando c, m' e inicializando con la matriz $U^{(0)} \in M_P$, de tal forma que los valores de la matriz de partición se actualicen por

$$\mu_{ik}^{(r+1)} = \left[\sum_{j=1}^{c} \left(\frac{d_{ik}^{(r)}}{d_{jk}^{(r)}} \right)^{\frac{2}{m'-1}} \right]^{-1}$$

hasta que la diferencia de matrices sea menor a un nivel de tolerancia ϵ, esto es

$$\left\| U^{(r+1)} - U^{r} \right\| \leq \epsilon.$$

1.1.3. Regresión logística

A fin de estudiar la dependencia entre una variable dependiente Y y más de una variable independiente $(x_1, x_2, \cdots, x_m)$, el modelo de regresión logístico, en el caso de que la variable dependiente sea dicotómica, está dado por:

$$P(Y = 1) = \frac{e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta \cdots + \beta_m x_m}}{1 + e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta \cdots + \beta_m x_m}}$$

Los parámetros de la ecuación de regresión logística se estiman por el método de máxima verosimilitud.

Johnson & Wichern (2002) discuten cómo esta regresión se puede usar para clasificar.

1.1.4. Análisis discriminante

Este método, propuesto por Fisher (1936), parte de n datos o individuos donde se han medido p variables cuantitativas independientes (explicativas) y una variable cualitativa (clasificativa) que posee las categorías de pertenencia, para obtener funciones lineales desde las variables explicativas que permitan comparar numéricamente un nuevo dato o individuo, al que no se le conoce la categoría. Este nuevo individuo pertenecerá al grupo que presente la función con mayor valor.

Dicha función, llamada discriminante de Fisher D, se obtiene como una función lineal de p variables explicativas

$$D = u_1X_1 + u_2X_2 + \cdots + u_pX_p$$

donde u_j corresponde a los coeficientes de ponderación, variando su valor de acuerdo con la categoría a la que pertenezca.

Considerando que el conjunto de entrenamiento cuenta con n individuos, se puede expresar la función discriminante, para cada uno, así:

$$D_i = u_1X_{1i} + u_2X_{2i} + \cdots + u_pX_{pi} \quad i = 1, 2, \cdots, n$$

donde D_i corresponde a la puntuación dicriminante de la i-ésima observación.

Puntuaciones discriminantes son los valores que resultan de evaluar cada individuo $(x_1, x_2, \cdots, x_p)$ en la función discriminante

$$D = u_1X_1 + u_2X_2 + \cdots + u_pX_p$$

que corresponde a la proyección de cada individuo sobre el eje discriminante.

Centroides son vectores de medias que resumen la información sobre los grupos. Los centroides de cada grupo están dados por:

$$\bar{X}_I = \begin{bmatrix} \bar{X}_{1,I} \\ \bar{X}_{2,I} \\ \vdots \\ \bar{X}_{p,I} \end{bmatrix} \quad \bar{X}_{II} = \begin{bmatrix} \bar{X}_{1,II} \\ \bar{X}_{2,II} \\ \vdots \\ \bar{X}_{p,II} \end{bmatrix}$$

Para los grupos I y II se tiene que la puntuación discriminante para el vector de medias, está dado por:

$$\bar{D}_I = u_1\bar{X}_{1,I} + u_2\bar{X}_{2,I} + \cdots + u_p\bar{X}_{p,I}$$
$$\bar{D}_{II} = u_1\bar{X}_{1,II} + u_2\bar{X}_{2,II} + \cdots + u_p\bar{X}_{p,II}$$

Punto de corte discriminante. Se calcula promediando la puntuación discriminante de los vectores de medias para cada grupo

$$C = \frac{\bar{D}_I + \bar{D}_{II}}{2}$$

Así, el criterio de clasificacón para el i-ésimo individuo será

$$\begin{cases} i \in I & \text{si} \quad D_i - C < 0 \\ i \in II & \text{si} \quad D_i - C > 0 \end{cases}$$

2. Objetivo del estudio

Determinar la eficiencia de los clasificadores SVM, LR, FC y LDA de acuerdo con sus FDR, para dos grupos de datos provenientes de algunas poblaciones multivariadas, con diferentes vectores de medias e identificar escenarios que permitan evaluar los méritos o debilidades de cada uno de ellos.

2.1. Metodología

Se compara la FDR, vía simulación, de las técnicas de clasificación SVM, FC, LDA y RL en tres tipos de distribuciones multivariadas como MSND (Azzalini & Dalla Va-lle 1996), MND (Anderson et al. 1958) y MTD (Kotz & Nadarajah 2004) en situaciones donde varía el tamaño de muestra y los parámetros para cada cantidad de variables.

2.2. Escenarios de simulación

Para los escenarios considerados en este estudio se tuvieron en cuenta los elementos comunes en la generación de dos grupos de datos, diferentes, provenientes de igual distribución; dichas distribuciones son MND, MSND y MTD. Además, se plantean, para estos conjuntos de datos, situaciones en las que varia el número de variables (2, 3 y 4), las distancias entre los vectores de medias (2, 3, 4, 5 y 6) y el tamaño de muestra (20, 50 y 100).

A continuación se expondrán los parámetros que componen cada escenario.

2.2.1. Escenario 1

En este escenario los grupos de datos son de dos variables con tamaños de muestra 20, 50 y 100, en los cuales permanece constante la matriz de varianzas y covarianzas dada por

$$\begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$$

Dicha matriz corresponde a la utilizada por diferentes autores en trabajos previos (Barajas & Morales 2009, Salazar et al. 2012).

Para la distribución MSND también permanece constante el parámetro de forma dado por $[-1, 1]$ y para la distribución MTD, también permanecen constantes los grados de libertad que en este caso son 10.

Se analizan cinco situaciones que comprenden el cambio en la distancia entre los vectores de medias. Así, se analizan los siguientes casos:

Caso 1: los vectores de medias para los dos grupos tienen una distancia de 2, los vectores de medias correspondientes son $[-1, 0]$ y $[1, 0]$.

Caso 2: los vectores de medias para los dos grupos tienen una distancia de 3, los vectores de medias correspondientes son $[-1, 0]$ y $[2, 0]$.

Caso 3: los vectores de medias para los dos grupos tienen una distancia de 4, los vectores de medias correspondientes son $[-2, 0]$ y $[2, 0]$.

Caso 4: los vectores de medias para los dos grupos tienen una distancia de 5, los vectores de medias correspondientes son $[-2, 0]$ y $[3, 0]$.

Caso 5: los vectores de medias para los dos grupos tienen una distancia de 6, los vectores de medias correspondientes son $[-3, 0]$ y $[3, 0]$.

2.2.2. Escenario 2

En este escenario los grupos de datos son de tres variables, con tamaños de muestra 20, 50 y 100, permaneciendo constante la matriz de varianzas y covarianzas, dada por

$$\begin{bmatrix} 1 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 1 \end{bmatrix}$$

La anterior matriz se toma con características similares a la matriz usada en el caso de datos con dos variables, es decir, la varianza de las variables es 1 y la covarianza entre pares de variables es 0.3.

Para la distribución MSND también permanece constante el parámetro de forma dado por $[-2, 0, 2]$, y para la distribución MTD permanecen constantes los grados de libertad que son 10.

Se presentan cinco situaciones en las cuales hay una distancia que va en aumento entre los vectores de medias de cada conjunto de datos analizado, esto es

Caso 1: los vectores de medias para los dos grupos tienen una distancia de 2, los vectores de medias correspondientes son $[-1, 0, 0]$ y $[1, 0, 0]$.

Caso 2: los vectores de medias para los dos grupos tienen una distancia de 3, los vectores de medias correspondientes son $[-1, 0, 0]$ y $[2, 0, 0]$.

Caso 3: los vectores de medias para los dos grupos tienen una distancia de 4, los vectores de medias correspondientes son $[-2, 0, 0]$ y $[2, 0, 0]$.

Caso 4: los vectores de medias para los dos grupos tienen una distancia de 5, los vectores de medias correspondientes son $[-2, 0, 0]$ y $[3, 0, 0]$.

Caso 5: los vectores de medias para los dos grupos tienen una distancia de 6, los vectores de medias correspondientes son $[-3, 0, 0]$ y $[3, 0, 0]$.

2.2.3. Escenario 3

En este escenario los grupos de datos son de cuatro variables, con tamaños de muestra 20, 50 y 100, permaneciendo constante la matriz de varianzas y covarianzas, dada por

$$\begin{bmatrix} 1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1 \end{bmatrix}$$

De igual forma, que el caso de los datos de tres variables, se toma la varianza de las variables como 1 y la covarianza como 0.3.

Para la distribución MSND también permanece constante el parámetro de forma dado por $[-2, -1, 1, 2]$, y para la distribución MTD permanecen constante los grados de libertad que son 10.

Se presentan cinco situaciones en las cuales hay una distancia que va en aumento entre los vectores de medias de cada conjunto de datos analizado, esto es

Caso 1: los vectores de medias para los dos grupos tienen una distancia de 2, los vectores de medias correspondientes son $[-1, 0, 0, 0]$ y $[1, 0, 0, 0]$.

Caso 2: los vectores de medias para los dos grupos tienen una distancia de 3, los vectores de medias correspondientes son $[-1, 0, 0, 0]$ y $[2, 0, 0, 0]$.

Caso 3: los vectores de medias para los dos grupos tienen una distancia de 4, los vectores de medias correspondientes son $[-2, 0, 0, 0]$ y $[2, 0, 0, 0]$.

Caso 4: los vectores de medias para los dos grupos tienen una distancia de 5, los vectores de medias correspondientes son $[-2, 0, 0, 0]$ y $[3, 0, 0, 0]$.

Caso 5: los vectores de medias para los dos grupos tienen una distancia de 6, los vectores de medias correspondientes son $[-3, 0, 0, 0]$ y $[3, 0, 0, 0]$.

2.2.4. Procedimiento de comparación

Las cuatro técnicas de clasificación, SVM, FC, LR y LD, fueron comparadas a partir de una secuencia de pasos, descritos a continuación:

Paso 1: se simularon dos conjuntos de datos de cada distribución, agrupados en un arreglo llamado *conjunto a clasificar*.

Paso 2: se simularon dos conjuntos de datos de cada distribución, con un tamaño del 50 % de los datos a clasificar, con idénticas características del conjunto de datos a clasificar.

Paso 3: con los datos simulados en el paso 2 se realizó el entrenamiento de los clasificadores SVM, LR y LD.

Paso 4: se clasificaron los datos obtenidos en el paso 1.

Paso 5: se calcularon las FDR para las clasificaciones obtenidas en el paso 4.

Paso 6: se repite 5000 veces del paso 1 al paso 5, luego se hace un promedio de todas las FDR obtenidas en estas 5000 repeticiones.

Todo el proceso de comparación que se realiza en este trabajo fue llevado a cabo por medio del paquete de uso público R. Para SVM se uso la función `svm()` de la librería e1071 del paquete class, para FC se uso la función `cmeans()` del paquete e1071, para LR se uso la función `glm()` del paquete stats y para LDA se usa la función `lda()` del paquete MASS.

2.3. Resultados

2.3.1. Escenario 1

En todos los casos considerados en este escenario se obtuvo que el clasificador que mejor se comporta es LDA seguido de LR, entre las FDR de estos dos clasificadores se conserva una diferencia entre 1 % y 2 %, como se muestra en la Figura 3.

Para las distribuciones MND y MTD se observa que subsiste, para todas las distancias entre medias consideradas, el comportamiento de ser el clasificador FC superior a SVM, caso contrario ocurre para la distribución MSND para la cual es más eficiente SVM que FC, esto observado a partir de las FDR calculadas en el proceso de simulación.

Se observa además que, cuando las distancias entre las medias de los distintos grupos de datos es mayor, para las distribuciones MND y MSND la tendencia de los clasificadores es comportarse de igual forma, es decir, obtienen similares FDR y cercanas a cero; mientras que para la distribución MTD se sigue comportando el patrón observado en la mínima distancia entre medias, incluso para esta distribución, en caso de tener mayor distancia entre medias, FC es más eficiente que LR.

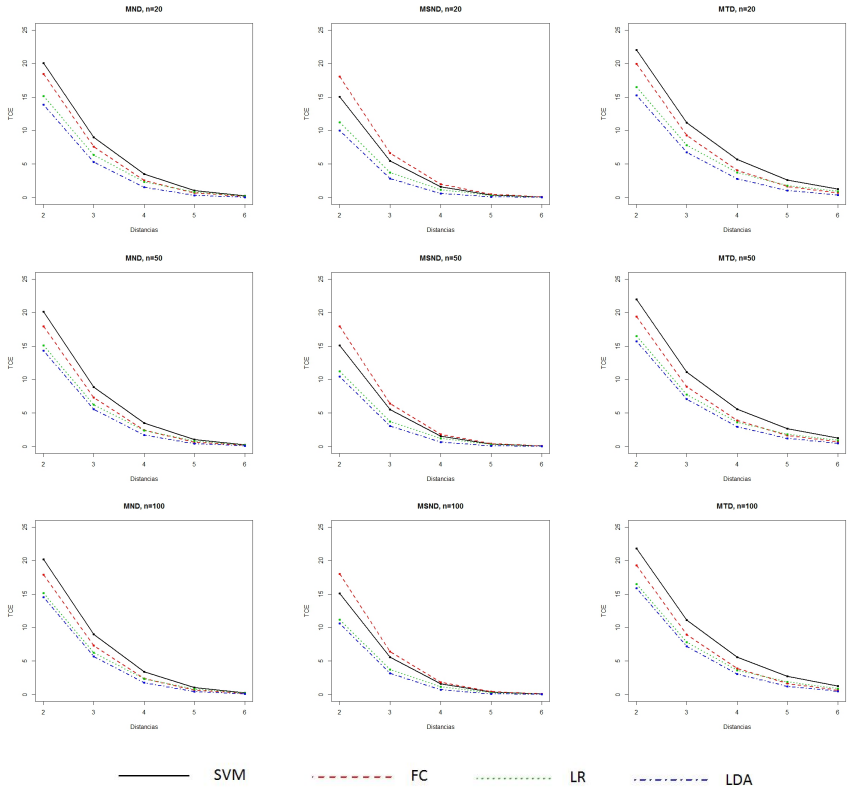

Figura 3: *resultados del escenario 1, con dos variables, con diferentes distancias entre medias e igual varianza. De arriba hacia abajo se cuentan los tamaños de muestra* 20, 50 *y* 100, *respectivamente, y de izquierda a derecha se observan las distribuciones MND, MSND y MTD, respectivamente. Fuente: elaboración propia.*

2.3.2. Escenario 2

Para los diferentes casos considerados en este escenario se observa que el clasificador LDA es el de mejor desempeño seguido de LR. Además, es de notar que en las distribuciones MND y MTD el clasificador FC tiene un mejor desempeño que el clasificador SVM sin importar la distancia entre las medias de los dos grupos de datos considerados.

Para la distribución MSND se observa que SVM posee un mejor desempeño que FC y que en el caso en que las distancias entre medias aumenta las FDR en esta distribución, en los clasificadores, tiende a comportarse igual que en la distribución MND obteniendo valores muy semejantes, caso contrario con la distribución MTD, que se alcanza a observar una leve diferencia entre la eficiencia de los clasificadores.

En la distribución MTD, cuando aumenta la distancia entre la media de los dos grupos de datos, mejora la eficiencia del clasificador FC, siendo superior a LR (*Figura 4*).

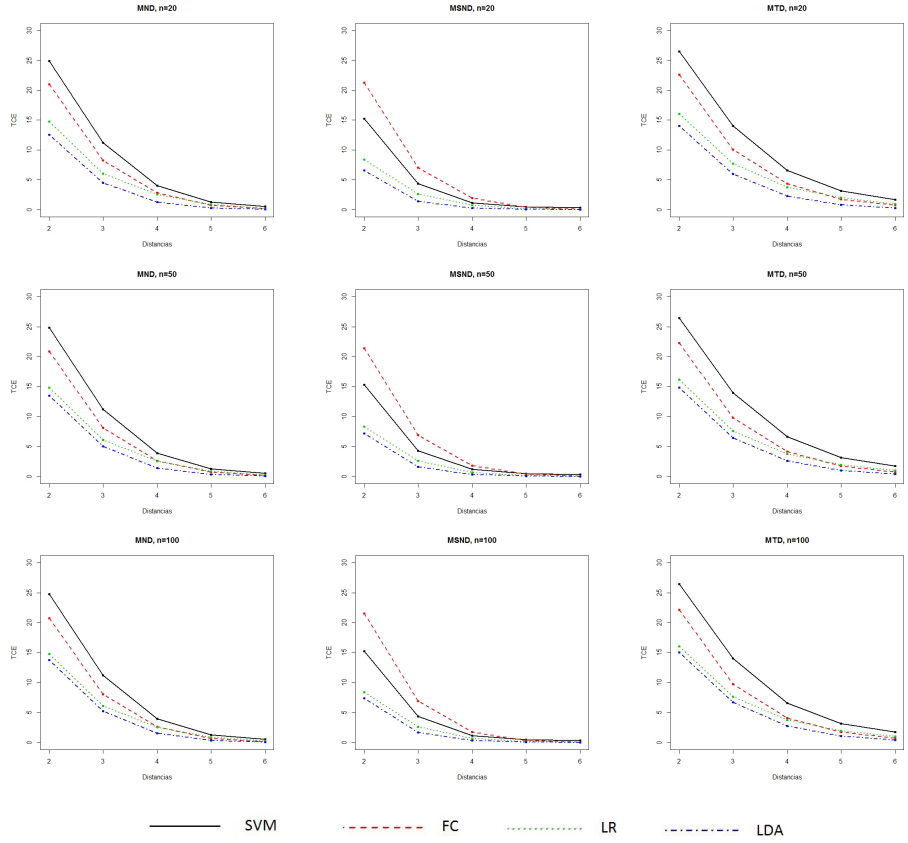

Figura 4: *resultados del escenario 2, con tres variables, con diferentes distancias entre medias e igual varianza. De arriba hacia abajo se cuentan los tamaños de muestra* 20, 50 *y* 100, *respectivamente, y de izquierda a derecha se observan las distribuciones MND, MSND y MTD, respectivamente. Fuente: elaboración propia.*

2.3.3. Escenario 3

En las distribuciones MND y MTD se presenta un mismo comportamiento entre las FDR de los clasificadores para los diferentes casos, esto es, para una distancia muy pequeña entre las medias de los grupos de datos, SVM y FC presentan igual comportamiento y a medida que aumenta dicha distancia FC presenta una mejor eficiencia que SVM y LR.

Para el caso de la distribución MSND se observa que a menor distancia entre medias es menor la eficiencia de FC respecto a SVM, pero cuando la distancia entre las medias supera las 4 unidades, FC supera en eficiencia a SVM.

Para los diferentes casos considerados se observa que el clasificador con mejor desempeño es LDA seguido de LR (ver figura 5).

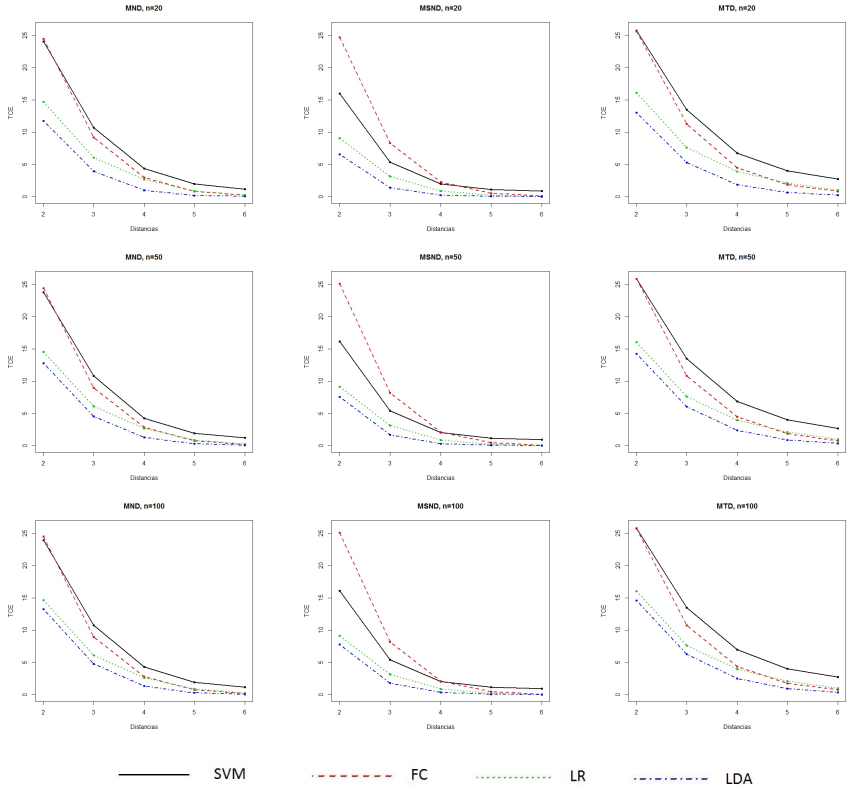

Figura 5: *se muestran los gráficos para el escenario 4, con cuatro variables, con diferentes distancias entre medias e igual varianza. De arriba hacia abajo se cuentan los tamaños de muestra* 20, 50 *y* 100*, respectivamente, y de izquierda a derecha se observan las distribuciones MND, MSND y MTD, respectivamente. Fuente: elaboración propia.*

3. Conclusiones

De los cuatro clasificadores analizados en este estudio, se encontró que LDA fue el de mejor desempeño seguido de LR, teniendo en cuenta las distribuciones consideradas.

Para los datos provenientes de una distribución MND se obtiene un mejor desempeño de FC con respecto a SVM cuando los datos poseen dos y tres variables, en el caso de cuatro variables este comportamiento es similar solo cuando las distancias ente las medias de los grupos aumenta.

Para emplear los clasificadores SVM, LR y LDA es necesario contar con un conjunto de datos de entrenamiento a fin de clasificar los datos en los que no se conoce su procedencia, por tal razón se hace importante la evaluación del clasificador FC ya que no necesita de conjunto de entrenamiento para hacer la clasificación.

Para los clasificadores considerados en los diferentes escenarios se puede encontrar que en cada distribución se trata de conservar un comportamiento de las FDR, a medida que aumenta el número de variables en los grupos de datos.

4. Agradecimientos

Agradecemos a la Escuela de Estadística de la Universidad Nacional de Colombia Sede Medellín.

Recibido: 07 de febrero de 2014
Aceptado: 09 de septiembre de 2014

Referencias

Anderson, T. W., Anderson, T. W., Anderson, T. W. & Anderson, T. W. (1958), *An introduction to multivariate statistical analysis*, Vol. 2, Wiley New York.

Azzalini, A. & Capitanio, A. (1999), 'Statistical applications of the multivariate skew normal distribution', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**(3), 579–602.

Azzalini, A. & Dalla Va-lle, A. (1996), 'The multivariate skew-normal distribution', *Biometrika* **83**(4), 715–726.

Azzalini, A. & Dalla Valle, A. (1996), 'The multivariate skew-normal distribution', *Biometrika* **83**(4), 715–726.

Barajas, F. H. & Morales, J. C. C. (2009), 'Comparación entre tres técnicas de clasificación', *Revista Colombiana de Estadística* **32**, 247–265.

Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Machine learning* **20**(3), 273–297.

Fisher, R. A. (1936), 'The use of multiple measurements in taxonomic problems', *Annals of eugenics* **7**(2), 179–188.

Hoppner, F., Klawonn, F., Kruse, R. & Runkler, T. (1999), *Fuzzy cluster analysis: methods for classification, data analysis and image recognition*, J. Wiley New York.

Hosmer, D. W. & Lemeshow, S. (2000), *Applied Logistic Regression (Wiley Series in Probability and Statistics)*, Wiley-Interscience Publication.

Johnson, R. A. & Wichern, D. W. (2002), *Applied multivariate statistical analysis*, Vol. 5, Prentice hall Upper Saddle River, NJ.

Kotz, S. & Nadarajah, S. (2004), *Multivariate T-Distributions and Their Applications*, Cambridge University Press.

Salazar, D. A., Vélez, J. I. & Salazar, J. C. (2012), 'Comparison between svm and logistic regression: Which one is better to discriminate?', *Revista Colombiana de Estadística* **35**(SPE2), 223–237.

A. Anexos

Las tablas presentadas a continuación muestran las TCE calculadas a los clasificadores estadísticos SVM, FC, LR y LDA en el proceso de simulación en escenarios donde se tienen en cuenta datos con 2, 3 y 4 variables provenientes de las distribuciones multivariadas MND, MSND y MTD y con parámetros establecidos previamente donde permanece constante la variabilidad entre los datos. Los gráficos relacionados con estas tablas corresponden a 3, 4 y 5. Los términos D1, D2, D3, D4 y D5 corresponden a las diferentes distancias entre los vectores de medias de los dos grupos considerados en cada caso de los escenarios; los respectivos valores son 2, 3, 4, 5 y 6.

Tabla 1: *TCE para datos de dos variables y tamaño de muestra* 20 *con la matriz previa. Fuente: elaboración propia.*

		SVM	FC	LR	LDA
MND	D1	20,079	18,435	15,164	13,879
	D2	8,9950	7,5825	6,3530	5,2800
	D3	3,4930	2,5810	2,3585	1,5605
	D4	1,0745	0,7090	0,8420	0,3305
	D5	0,2770	0,1575	0,2235	0,0645
MSND	D1	15,0230	18,0740	11,2115	10,0160
	D2	5,4725	6,6660	3,7125	2,8415
	D3	1,6095	1,9875	1,1870	0,6170
	D4	0,381	0,482	0,293	0,103
	D5	0,1165	0,0910	0,0430	0,0125
MTD	D1	22,0245	19,9455	16,5220	15,2820
	D2	11,1840	9,2870	7,8005	6,7820
	D3	5,6855	4,0460	3,7140	2,7915
	D4	2,6235	1,6760	1,7875	1,0280
	D5	1,2840	0,7265	0,9105	0,3965

Tabla 2: *TCE para datos de dos variables y tamaño de muestra* 50 *con la matriz previa. Fuente: elaboración propia.*

		SVM	FC	LR	LDA
MND	D1	20,1274	17,9484	15,1128	14,3226
	D2	8,8966	7,3166	6,2388	5,5744
	D3	3,4902	2,4624	2,4032	1,7140
	D4	1,0534	0,6696	0,8440	0,4108
	D5	0,2884	0,1486	0,2244	0,0696
MSND	D1	15,0968	17,9792	11,2146	10,4674
	D2	5,5404	6,4450	3,7550	3,0746
	D3	1,5608	1,8194	1,2022	0,6428
	D4	0,3738	0,4338	0,2920	0,1108
	D5	0,1122	0,0730	0,0510	0,0148
MTD	D1	21,9566	19,3858	16,4854	15,6854
	D2	11,1340	8,9340	7,7882	7,0844
	D3	5,5796	3,8912	3,6334	2,9464
	D4	2,6836	1,6424	1,8136	1,1990
	D5	1,2780	0,7060	0,9102	0,4792

Tabla 3: *TCE para datos de dos variables y tamaño de muestra* 100 *con la matriz previa. Fuente: elaboración propia.*

		SVM	FC	LR	LDA
MND	D1	20,1783	17,9231	15,1273	14,5297
	D2	8,9752	7,3165	6,2629	5,6985
	D3	3,4013	2,4134	2,3466	1,7550
	D4	1,0486	0,6456	0,8203	0,4055
	D5	0,2915	0,1370	0,2181	0,0766
MSND	D1	15,1057	18,0197	11,1976	10,6089
	D2	5,5710	6,3992	3,7365	3,1474
	D3	1,5956	1,8349	1,1844	0,6937
	D4	0,3684	0,4189	0,2877	0,1117
	D5	0,1219	0,0740	0,0479	0,0145
MTD	D1	21,8227	19,3089	16,5112	15,8854
	D2	11,1444	8,9236	7,8442	7,2279
	D3	5,5844	3,8767	3,6493	3,0367
	D4	2,7062	1,6451	1,8627	1,2413
	D5	1,2738	0,6807	0,8941	0,4993

Tabla 4: *TCE para datos de tres variables y tamaño de muestra* 20 *con la matriz previa. Fuente: elaboración propia.*

		SVM	FC	LR	LDA
MND	D1	24,9315	21,0350	14,7265	12,5620
	D2	11,1935	8,2145	6,0550	4,4805
	D3	3,9815	2,8150	2,5305	1,2625
	D4	1,2570	0,7270	0,8645	0,2220
	D5	0,5355	0,1715	0,2080	0,0320
MSND	D1	15,2130	21,2715	8,3610	6,5955
	D2	4,3220	7,0495	2,5810	1,4140
	D3	1,1330	1,9115	0,6920	0,2325
	D4	0,4555	0,4005	0,1295	0,0345
	D5	0,3180	0,0640	0,0220	0,0055
MTD	D1	26,4795	22,5905	16,0610	13,9875
	D2	14,0295	10,0645	7,6820	5,9885
	D3	6,5835	4,3240	3,7625	2,2460
	D4	3,168	1,735	1,966	0,792
	D5	1,6755	0,7095	0,9325	0,2825

Tabla 5: *TCE para datos de tres variables y tamaño de muestra* 50 *con la matriz previa. Fuente: elaboración propia.*

		SVM	FC	LR	LDA
MND	D1	24,8266	20,8452	14,8042	13,4994
	D2	11,1986	8,0846	6,1038	4,9872
	D3	3,8800	2,6238	2,5088	1,4088
	D4	1,2890	0,7044	0,8768	0,3062
	D5	0,5430	0,1494	0,2194	0,0490
MSND	D1	15,3070	21,4084	8,3322	7,1862
	D2	4,2482	6,9060	2,5686	1,5608
	D3	1,1700	1,7690	0,6756	0,3004
	D4	0,4608	0,3536	0,1206	0,0442
	D5	0,3134	0,0608	0,0172	0,0048
MTD	D1	26,4140	22,2752	16,1380	14,8408
	D2	13,9442	9,7958	7,6004	6,4420
	D3	6,6060	4,1144	3,7628	2,6024
	D4	3,1270	1,7178	1,9418	0,9742
	D5	1,7444	0,7200	0,9620	0,3836

Tabla 6: *TCE para datos de tres variables y tamaño de muestra* 100 *con la matriz previa. Fuente: elaboración propia.*

		SVM	FC	LR	LDA
MND	D1	24,7795	20,7427	14,7708	13,7664
	D2	11,2212	8,0299	6,1119	5,1890
	D3	3,9368	2,6172	2,5007	1,5105
	D4	1,2874	0,6836	0,8691	0,3307
	D5	0,5275	0,1459	0,2285	0,0548
MSND	D1	15,2221	21,5278	8,3441	7,4012
	D2	4,3216	6,9083	2,5867	1,6577
	D3	1,1351	1,7264	0,6730	0,3111
	D4	0,4756	0,3457	0,1174	0,0535
	D5	0,3216	0,0531	0,0187	0,0071
MTD	D1	26,4693	22,1564	16,0423	15,0206
	D2	14,0146	9,6894	7,6410	6,6965
	D3	6,5530	4,0622	3,7390	2,7092
	D4	3,1558	1,7143	1,9602	1,0681
	D5	1,7429	0,7095	0,9553	0,4163

Tabla 7: *TCE para datos de cuatro variables y tamaño de muestra* 20 *con la matriz previa. Fuente: elaboración propia.*

		SVM	FC	LR	LDA
MND	D1	24,0760	24,4910	14,7365	11,7615
	D2	10,7210	9,1750	6,0630	3,345
	D3	4,3270	2,9480	2,6830	1,0015
	D4	1,9730	0,7960	0,8855	0,1730
	D5	1,1350	0,1895	0,2335	0,0215
MSND	D1	15,9805	24,7085	9,0485	6,5650
	D2	5,3640	8,3260	3,1505	1,3850
	D3	1,9785	2,2290	0,8620	0,2130
	D4	1,1065	0,4980	0,1705	0,0350
	D5	0,8605	0,0835	0,0295	0,0020
MTD	D1	25,7465	25,6810	16,1195	13,0370
	D2	13,4720	11,2890	7,6005	5,2695
	D3	6,7185	4,4585	3,8725	1,8505
	D4	3,9875	1,8560	2,0675	0,6445
	D5	2,7430	0,7795	1,0010	0,1920

Tabla 8: *TCE para datos de cuatro variables y tamaño de muestra* 50 *con la matriz previa. Fuente: elaboración propia.*

		SVM	FC	LR	LDA
MND	D1	23,8172	24,4402	14,5366	12,7764
	D2	10,8030	8,9586	6,0838	4,5452
	D3	4,2286	2,8234	2,6370	1,2432
	D4	1,9286	0,7600	0,8754	0,2546
	D5	1,1882	0,1548	0,2072	0,0362
MSND	D1	16,1442	25,1298	9,1520	7,5222
	D2	5,4064	8,1862	3,1186	1,6570
	D3	2,0152	2,0920	0,8692	0,2940
	D4	1,1334	0,4652	0,1804	0,0558
	D5	0,8874	0,0664	0,0282	0,0066
MTD	D1	25,8752	25,9002	16,0256	14,2242
	D2	13,4882	10,8236	7,6100	6,0208
	D3	6,8780	4,4424	3,9474	2,3794
	D4	4,0058	1,8178	2,0960	0,8806
	D5	2,6832	0,7442	0,9836	0,3052

Tabla 9: *TCE para datos de cuatro variables y tamaño de muestra* 100 *con la matriz previa. Fuente: elaboración propia.*

		SVM	FC	LR	LDA
MND	D1	23,9492	24,5697	14,6550	13,2641
	D2	10,7729	8,9495	6,0914	4,7707
	D3	4,2849	2,7771	2,6210	1,3359
	D4	1,9033	0,7246	0,8857	0,2776
	D5	1,1741	0,1503	0,2189	0,0407
MSND	D1	16,0846	25,1375	9,1070	7,7995
	D2	5,3984	8,1875	3,1283	1,7937
	D3	2,0190	2,0705	0,8663	0,3341
	D4	1,1280	0,4422	0,1737	0,0577
	D5	0,8945	0,0690	0,0268	0,0066
MTD	D1	25,8418	25,7891	16,0764	14,5832
	D2	13,4752	10,7565	7,6313	6,2851
	D3	6,9418	4,3458	3,9415	2,5029
	D4	3,9907	1,7484	2,0626	0,9422
	D5	2,7007	0,7249	0,9879	0,3523

Tabla 10: *TCE para datos de dos variables y tamaño de muestra* 20 *con la matriz arbitraria. Fuente: elaboración propia.*

		SVM	FC	LR	LDA
MND	D1	41,3610	37,9550	34,1205	31,8455
	D2	33,7290	33,1320	26,5110	24,7775
	D3	26,078	27,882	20,154	18,702
	D4	19,3020	22,9035	14,6790	13,3565
	D5	14,1415	18,5870	10,3845	9,1755
MSND	D1	39,2365	37,9560	31,7870	29,9320
	D2	30,2875	32,9335	23,8495	22,2670
	D3	22,2120	27,7160	17,2275	15,7990
	D4	15,7735	22,8330	11,8800	10,6500
	D5	10,8380	18,1725	7,7515	6,6980
MTD	D1	41,8415	38,3335	34,6820	32,4885
	D2	34,303	33,840	27,353	25,689
	D3	27,3270	28,7600	21,1935	19,9005
	D4	21,0605	24,0325	16,0560	14,8410
	D5	16,0470	19,8395	11,8725	10,6845

Tabla 11: *TCE para datos de dos variables y tamaño de muestra* 50 *con la matriz arbitraria. Fuente: elaboración propia.*

		SVM	FC	LR	LDA
MND	D1	41,6466	38,6090	34,1176	32,9644
	D2	33,8258	33,0050	26,5192	25,5926
	D3	25,8726	27,7568	20,0706	19,2418
	D4	19,3806	22,8600	14,7634	14,0044
	D5	14,1758	18,2842	10,5010	9,7532
MSND	D1	39,1782	38,5552	31,8758	30,7582
	D2	30,2760	32,9698	23,9978	23,0818
	D3	22,0686	27,5854	16,9986	16,1778
	D4	15,6136	22,4484	11,7204	10,9744
	D5	10,9628	18,0224	7,8172	7,1196
MTD	D1	41,7052	38,8736	34,6692	33,3922
	D2	34,4924	33,7318	27,4408	26,4796
	D3	27,1954	28,5472	21,1710	20,3916
	D4	21,0960	23,8294	16,1190	15,2982
	D5	16,1110	19,4238	11,9764	11,2356

Tabla 12: *TCE para datos de dos variables y tamaño de muestra* 100 *con la matriz arbitraria. Fuente: elaboración propia.*

		SVM	FC	LR	LDA
MND	D1	41,5540	38,6178	34,1229	33,2333
	D2	33,8096	33,1090	26,6035	25,8573
	D3	26,0888	27,7798	20,0922	19,4575
	D4	19,3548	22,7141	14,6753	14,0797
	D5	14,0968	18,2712	10,4531	9,8832
MSND	D1	39,4210	38,5090	32,0009	31,1496
	D2	30,3461	33,0264	23,8794	23,1730
	D3	22,2203	27,5318	17,1018	16,4760
	D4	15,6959	22,4741	11,7604	11,1711
	D5	10,8189	17,8785	7,7375	7,2013
MTD	D1	41,8471	39,0411	34,6854	33,7403
	D2	34,3024	33,7041	27,3592	26,6112
	D3	27,2402	28,5509	21,2455	20,5162
	D4	21,1519	23,7124	16,0580	15,4169
	D5	16,2035	19,3876	11,8724	11,2963

Tabla 13: *TCE para datos de tres variables y tamaño de muestra* 20 *con la matriz arbitraria. Fuente: elaboración propia.*

		SVM	FC	LR	LDA
MND	D1	26,3235	38,8400	16,4160	14,1860
	D2	13,5620	32,9150	7,4755	5,8380
	D3	5,8835	25,4895	3,2190	1,8295
	D4	2,1505	17,1160	1,3160	0,4500
	D5	0,8260	10,7310	0,3915	0,0955
MSND	D1	21,2415	38,8815	13,0020	10,8855
	D2	9,0595	33,0405	4,9070	3,3975
	D3	3,0870	25,2285	1,8750	0,7760
	D4	1,0435	17,0250	0,5620	0,1400
	D5	0,4345	10,0850	0,1190	0,0200
MTD	D1	27,6785	39,1105	18,0550	15,7205
	D2	15,9265	33,7700	8,9100	7,1650
	D3	8,3610	26,9035	4,5190	2,9735
	D4	4,2775	19,3605	2,4665	1,1750
	D5	2,2365	12,7405	1,2940	0,4260

Tabla 14: *TCE para datos de tres variables y tamaño de muestra* 50 *con la matriz arbitraria. Fuente: elaboración propia.*

		SVM	FC	LR	LDA
MND	D1	26,3832	39,4604	16,4972	15,2064
	D2	13,6818	33,1518	7,4442	6,3258
	D3	5,7848	25,3572	3,1632	2,1036
	D4	2,1380	17,4054	1,3110	0,5560
	D5	0,8312	10,7278	0,4230	0,1062
MSND	D1	21,2056	39,6542	12,8496	11,6806
	D2	9,1108	33,4354	4,8796	3,8234
	D3	3,0640	25,4018	1,8748	0,9526
	D4	1,0028	17,0372	0,5498	0,1762
	D5	0,4532	10,3068	0,1266	0,0246
MTD	D1	27,6636	39,8602	17,9286	16,5638
	D2	15,9890	34,1330	9,0126	7,8560
	D3	8,3970	26,8454	4,4976	3,3752
	D4	4,2388	19,3926	2,5056	1,4288
	D5	2,2688	12,9422	1,3104	0,5618

Tabla 15: *TCE para datos de tres variables y tamaño de muestra* 100 *con la matriz arbitraria. Fuente: elaboración propia.*

		SVM	FC	LR	LDA
MND	D1	26,5212	39,6513	16,5865	15,5764
	D2	13,5361	33,1972	7,4105	6,4806
	D3	5,7367	25,3820	3,1752	2,1732
	D4	2,1055	17,3168	1,3091	0,5692
	D5	0,8388	10,7430	0,4006	0,1162
MSND	D1	21,1302	39,8615	12,8740	11,9169
	D2	9,0729	33,3219	4,8795	3,9317
	D3	3,0714	25,4332	1,8927	1,0076
	D4	1,0089	17,0478	0,5582	0,1958
	D5	0,4618	10,3443	0,1265	0,0319
MTD	D1	27,5503	40,0627	17,8492	16,7743
	D2	15,8591	34,0000	8,9831	8,0243
	D3	8,4094	26,7743	4,5339	3,5460
	D4	4,2138	19,2877	2,4772	1,4960
	D5	2,2283	12,8423	1,3247	0,6184

Tabla 16: *TCE para datos de cuatro variables y tamaño de muestra* 20 *con la matriz arbitraria. Fuente: elaboración propia.*

		SVM	FC	LR	LDA
MND	D1	28,8025	39,1100	16,9260	13,7180
	D2	15,0475	33,6260	7,7325	5,4725
	D3	6,7000	26,1565	3,5510	1,6210
	D4	2,9040	18,0645	1,5030	0,4120
	D5	1,4185	11,2195	0,4790	0,0835
MSND	D1	24,7055	39,2555	14,4180	11,4875
	D2	11,0955	33,5885	6,0570	3,8980
	D3	4,2010	25,5910	2,6515	0,9770
	D4	1,6690	16,9550	0,9025	0,1870
	D5	0,8695	9,6985	0,2275	0,0280
MTD	D1	29,6805	39,5675	18,1320	15,0220
	D2	17,3065	34,4740	9,2925	6,7890
	D3	9,4090	27,6880	4,8975	2,7645
	D4	5,0540	20,0865	2,8355	1,0745
	D5	3,0385	13,5085	1,4855	0,3725

Tabla 17: *TCE para datos de cuatro variables y tamaño de muestra* 50 *con la matriz arbitraria. Fuente: elaboración propia.*

		SVM	FC	LR	LDA
MND	D1	28,7308	40,1916	16,8858	15,0638
	D2	15,1292	34,0980	7,6596	6,1594
	D3	6,7448	26,4780	3,5872	2,0418
	D4	2,8832	18,1250	1,5154	0,5462
	D5	1,3718	11,3132	0,4732	0,1084
MSND	D1	24,7106	40,2210	14,4298	12,6656
	D2	11,2098	34,0482	6,0204	4,4914
	D3	4,2390	25,9024	2,5796	1,2122
	D4	1,6438	16,8174	0,8784	0,2510
	D5	0,8510	9,8310	0,2248	0,0462
MTD	D1	29,7500	40,3268	18,1868	16,2016
	D2	17,3318	34,7750	9,2544	7,6394
	D3	9,3066	27,8012	4,8820	3,2852
	D4	5,0680	20,1622	2,7468	1,3278
	D5	3,0374	13,4332	1,4582	0,5138

Tabla 18: *TCE para datos de cuatro variables y tamaño de muestra* 100 *con la matriz arbitraria. Fuente: elaboración propia.*

		SVM	FC	LR	LDA
MND	D1	28,6976	40,1997	16,8362	15,4205
	D2	15,0957	34,0215	7,6870	6,4014
	D3	6,7713	26,3989	3,5636	2,1729
	D4	2,8555	18,1747	1,4848	0,5802
	D5	1,3791	11,2252	0,4790	0,1201
MSND	D1	24,7162	40,3218	14,4226	13,0359
	D2	11,1250	34,0081	6,0113	4,6843
	D3	4,1945	25,9516	2,6478	1,3088
	D4	1,6631	17,0513	0,9100	0,2897
	D5	0,8545	9,8647	0,2382	0,0488
MTD	D1	29,8536	40,5036	18,1466	16,6762
	D2	17,3286	34,7146	9,2546	7,9199
	D3	9,3233	27,6491	4,9365	3,4775
	D4	5,0344	20,1332	2,7445	1,4472
	D5	3,0465	13,4219	1,4747	0,5860