Clasificación Supervisada LDA: Un Enfoque Robusto y No Paramétrico

Juan F. Arias[1], Juan P. Restrepo[1], Santiago Ortiz[2], Henry Laniado[2]

[1]Departamento de Ingeniería de Procesos, Universidad EAFIT; [2]Departamento de Ciencias Matemáticas, Universidad EAFIT
jariasa3@eafit.edu.co - jurest82@eafit.edu.co - sortiza2@eafit.edu.co - hlaniado@eafit.edu.co

Introducción

El Análisis Discriminante Lineal o LDA, es una técnica de clasificación supervisada en la que a partir de información de poblaciones conocidas, se clasifican otras observaciones en relación a sus propiedades o características (James et al., 2013). LDA utiliza la distancia de Mahalanobis para asignar un individuo en una de las poblaciones consideradas (Peña, 2002). Sin embargo, esta técnica es vulnerable a datos atípicos, en particular, cuando los parámetros considerados son estimados por máxima verosimilitud (Rousseeuw & Van Zomeren, 1990). La falta de robustez de este tipo de estimadores es una desventaja que ha llevado a desarrollar métodos que permitan tratar la presencia de datos atípicos. En esta investigación se presenta una alternativa robusta y no paramétrica del LDA. La propuesta se centra en estimar la estructura de covarianza como el producto entre una estimación no paramétrica de la correlación y una estimación robusta de las desviaciones estándar y el parámetro de localización a partir de las medianas marginales.

Método Propuesto

Sea $X_1 = [A, B]'$ una muestra aleatoria p–dimensional contaminada de la forma $(1-\alpha)A + \alpha B$ con $\alpha \in \left(0, \frac{1}{2}\right)$. $A \sim \mathcal{N}_p(\mu_A, \Sigma_A)$ y $B \sim \mathcal{N}_p(\mu_B, I)$, donde $||\mu_A - \mu_B||_2 = K$ para $K \gg 0$ y sea X_2 otra muestra p–dimensional, tal que $X_2 \sim \mathcal{N}_p(\mu_2, \Sigma_2)$. De acuerdo al LDA, la clasificiación de un nuevo individuo $Z \in \mathbb{R}^p$ viene dada por la comparación de las distancias de Mahalanobis al cuadrado (MD^2) de Z al centroide de cada población, es decir, $MD^2(Z, \mu_1)$ y $MD^2(Z, \mu_2)$ de la siguiente manera:

$$(Z - \hat{\mu}_1)' \, \hat{\Sigma}_1^{-1} \, (Z - \hat{\mu}_1) < (Z - \hat{\mu}_2)' \, \hat{\Sigma}_2^{-1} \, (Z - \hat{\mu}_2) \qquad (1)$$

Si (1) se cumple, entonces $Z \in X_1$, de los contrario, $Z \in X_2$. Es importante resaltar que, $\hat{\mu}_L$ y $\hat{\Sigma}_L$, para $L = 1, 2$, corresponden al vector de medias y a la matriz de covarianzas muestrales tanto para X_1 como para X_2. Dado que estos estimadores se afectan por la presencia de datos atípicos (Peña, 2002), se propone introducir nuevas versiones tanto de $\hat{\mu}_L$ como de $\hat{\Sigma}_L$:

$$\hat{\mu}_L^* = median(X_L)$$
$$\hat{\Sigma}_L^* = H_L \, \tau_{K_L} \, H_L$$

Donde τ_{K_L} corresponde a la matriz de correlación de Kendall de la L–ésima muestra y H_L es una matriz diagonal de orden p, cuyas entradas corresponden a los estimadores S_n por variable:

$$H_L = \begin{bmatrix} S_{n_{11}}(X_L) & & \\ & \cdots & \\ & & S_{n_{pp}}(X_L) \end{bmatrix}$$

Siendo $S_n(X) = median_i \left(median_j \left(|x_i - x_j|\right)\right)$ una altenativa robusta para la estimación de la desviación estándar (Rousseeuw & Croux, 1993). Este producto de matrices da como resultado un $\hat{\Sigma}^*$ con propiedades robustas y no–paramétricas.

Resultados

La Figura 1 (a) corresponde a las matrices de confusión de los métodos estándar y propuesto, al evaluarlos en un escenario donde se posicionaron las poblaciones de acuerdo a lo descrito en la metodología con un $\alpha = 0{,}3$. La Figura 1 (b) es el histograma de la Tasa de Falsos Positivos para 1000 movimientos aleatorios del vector de localización de la contaminación y de su proporción $\alpha \in [\frac{1}{10}, \frac{4}{10}]$. La Figura 1 (c) es la exactitud de ambos clasificadores a lo largo 1000 desplazamientos del vector de localización de B, aumentando la distancia euclidia del mismo respecto a μ_A.

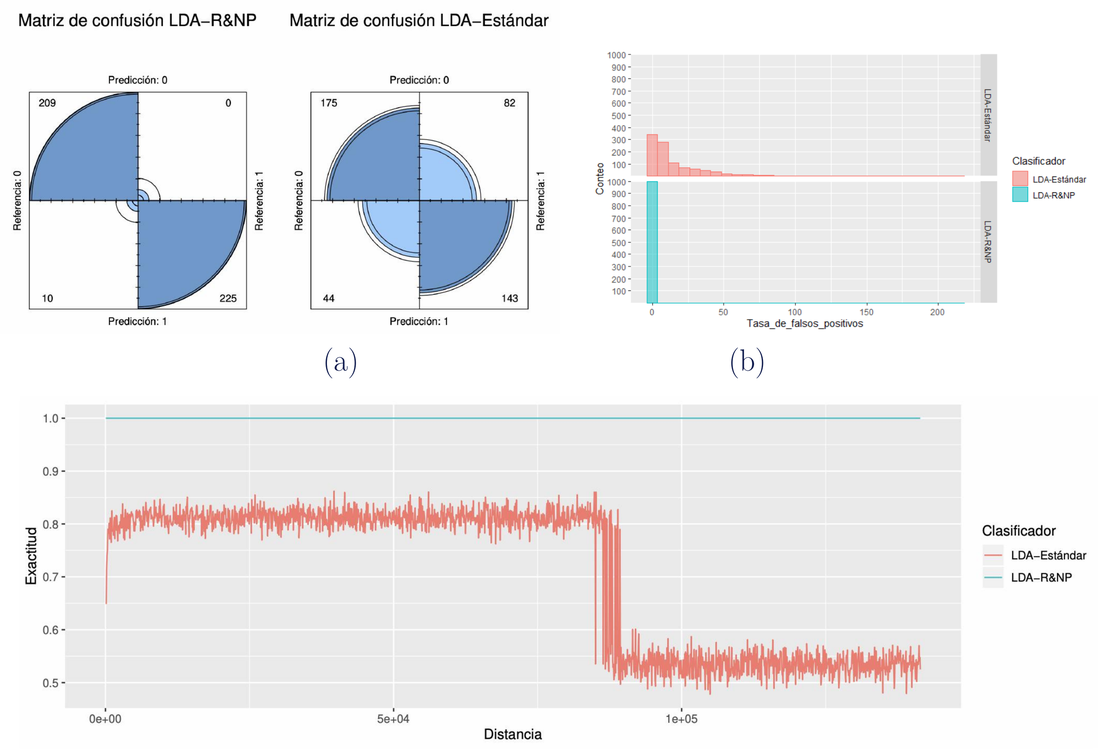

Figura 1: (a) Matrices de confusión del LDA RNP vs. LDA-Estándar, (b) Tasa de falsos positivos del LDA R&NP vs. LDA-Estándar a lo largo de 15 movimientos con variación de la contaminación, (c) Exactitud de ambos métodos respecto a la distancia euclidea de la contaminación.

Conclusiones

Se ha presentado una mejora a la técnica LDA estándar a partir de emplear estimadores robustos y no paramétricos en el cálculo de las estructuras de variabilidad y de locación. A partir de los resultados, se tiene una evidencia empírica de que el LDA-R&NP propuesto presenta un desempeño superior al del LDA-Estándar, por lo que representa una alternativa más eficiente en clasificación. Adicionalmente, es importante notar que el LDA-R&NP es consistente ante variaciones de la ubicación y la cantidad de contaminación, cualidad que no posee su contraparte estándar.

Referencias

1. James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.
2. Peña, D. (2002). Análisis de Datos Multivariantes. McGraw-Hill.
3. Rousseeuw, P. J. & Van Zomeren, B. C. (1990). Unmasking Multivariate Outliers and Leverage Points. Journal of the American Statistical Association, 85(411), 633-639.
4. Rousseeuw, P. J. & Croux, C. (1993). Alternatives to the Median Absolute Deviation. Journal of the American Statistical Association, 88(424), 1273-1283.