Una evaluación del desempeño en la clasificación binaria mediante simulación: Árboles de clasificación y Bosques aleatorios

Celina Beltrán; Ivana Barbona

Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Argentina

cbeltran2510@gmail.com

Abstract

This paper proposes the study of these multivariate statistical techniques: Classification Trees and Random Forests, being of interest to evaluate their performance when they are used in data that differ in the structure of correlations between the intervening variables and the sample sizes. A tool to evaluate these performances is through simulation. Four scenarios were defined for data simulation with the following characteristics:

- Scenario 1: Response variable highly correlated with the predictors and predictor variables little correlated with each other.
- Scenario 2: Response variable poorly correlated with the predictors and predictor variables highly correlated with each other.
- Scenario 3: Response variable highly correlated with the predictors and the predictor variables also highly correlated with each other.
- Scenario 4: Response variable little correlated with the predictors and likewise the predictor variables little correlated with each other.

In scenarios 1 and 3, the proposed situation corresponds to “separable” groups based on the values of the predictors; while in scenarios 2 and 4 the groups overlap with respect to the predictor variables, making it difficult to discriminate them based on the predictor variables. 500 data files were generated by simulation for each of the following sample sizes: 30, 75, 200, 400, 600, 1000. 20% of the observations were "marked" to be used as a test group and the rest 80% for the estimation of the models evaluated in each case.

As a main result, it is evident that, in those scenarios favorable to the classification by the structure of correlations of the variables that separate the groups (Scenarios 1 and 3), the evidence in favor of the Random Forest technique is significant, regardless of the sample

size. However, in the cases in which the response variable was not correlated with the explanatory ones, and therefore the groups are not able to be discriminated by the values of explanatory variables (Scenarios 2 and 4), there is no evidence of superiority of Forest technique except in isolated cases. This behavior of the Random Forests agrees with what is observed when evaluating the average percentage error of the forest according to the number of estimated trees. It is possible to distinguish different behaviors depending on the scenario. In the most favorable scenarios for the classification (Scenarios 1 and 3), the mean percentage error clearly decreases as the size of the forest and the sample size increase; while in cases where the group separation is not achieved by the explanatory variables, the mean percentage error seems to remain constant without showing an advantage regardless of the size of the forest and the data set. These results shed some light when choosing the appropriate statistical technique to classify units when the variables under consideration are or are not correlated and the response groups are overlapped or not with respect to their values.

Keywords: random forest; classification trees; simulation

Resumen

En este trabajo se propone el estudio de estas las técnicas estadísticas multivariadas Árboles de clasificación y Bosques aleatorios siendo de interés evaluar el desempeño de las mismas cuando son utilizadas en datos que difieren en la estructura de correlaciones entre las variables intervinientes y los tamaños de muestras. Una herramienta para evaluar estos desempeños es mediante simulación. Se definieron 4 escenarios para la simulación de datos con las siguientes características:

- Escenario 1: Variable respuesta altamente correlacionada con las predictoras y las variables predictoras poco correlacionadas entre sí.
- Escenario 2: Variable respuesta poco correlacionada con las predictoras y las variables predictoras muy correlacionadas entre sí.
- Escenario 3: Variable respuesta muy correlacionada con las predictoras y las variables predictoras también muy correlacionadas entre sí.
- Escenario 4: Variable respuesta poco correlacionada con las predictoras y asimismo las variables predictoras poco correlacionadas entre sí.

En los escenarios 1 y 3, la situación planteada se corresponde con grupos “separables” en función de los valores de los predictores; mientras que en los escenarios 2 y 4 los grupos están solapados respecto a las variables predictoras, dificultando la tarea de discriminarlos en función de las mismas. Se generaron mediante simulación 500 archivos de datos para cada uno de los siguientes tamaños de muestra: 30, 75, 200, 400, 600, 1000. Se “marcó” el 20% de las observaciones para ser utilizadas como grupo de test y el restante 80% para la estimación de los modelos evaluados en cada caso.

Como resultado principal se evidencia que, en aquellos escenarios donde es favorable la clasificación por la estructura de correlaciones de las variables que suponen una separación de los grupos (Escenarios 1 y 3), la evidencia en favor de la técnica de Bosques Aleatorios es significativa, independientemente del tamaño de muestra. Sin embargo, en los casos en que la variable respuesta no fue generada correlacionada con las explicativas, y por lo tanto los grupos no son capaces de ser discriminados por los valores de dichas variables (Escenarios 2 y 4), no hay evidencia de superioridad de la técnica de Bosques excepto en aislados casos. Este comportamiento de los Bosques Aleatorio concuerda con lo que se observa al evaluar el error medio porcentual del bosque según el número de árboles estimados. Es posible distinguir comportamientos diferentes según escenario. En los escenarios más favorables para la clasificación (Escenarios 1 y 3) el error medio porcentual disminuye claramente al incrementar el tamaño del bosque y el tamaño de muestra; mientras que en casos donde la separación de grupos no es lograda por las variables explicativas, el error porcentual medio parece mantenerse constante sin mostrar una ventaja independientemente del tamaño del bosque y del conjunto de datos. Estos resultados ponen cierta luz al momento de elegir la técnica estadística conveniente para clasificar unidades cuando las variables en consideración están o no correlacionadas y los grupos respuesta se encuentran solapados o no respecto a los valores de las mismas.

Palabras clave: redes neuronales; árboles de clasificación; simulación

1. Introducción

El Análisis Multivariado se refiere al tipo de análisis que se realiza sobre n unidades experimentales sobre las cuales se han medido p variables y se pretende estudiar a todas las variables (o un gran número) en forma simultánea (Hair *et al*., 1999). Uno de los objetivos de dichas técnicas es la clasificación de unidades en grupos que puede verse como la predicción de una variable categórica o cualitativa (Hastie *et al*., J. 2009). En la clasificación supervisada se cuenta con un conocimiento a priori, es decir para la tarea de clasificar un objeto dentro de una categoría o clase se cuenta con la información de p variables observadas en un conjunto de objetos cuya categoría o clase de pertenencia se conoce. Entre las técnicas de clasificación correspondiente al enfoque de minería de datos, se pueden citar: Árboles de clasificación y Bosques aleatorios. En este trabajo se propone el estudio de estas dos técnicas estadísticas multivariadas siendo de interés evaluar el desempeño de las mismas cuando son utilizadas en datos que difieren en la estructura de correlaciones entre las variables intervinientes y los tamaños de muestras. Una herramienta para evaluar estos desempeños es mediante simulación. Podríamos definir la simulación como una técnica que consiste en realizar experimentos de muestreo sobre el modelo de un sistema, con el objetivo de recopilar información bajo determinadas condiciones. Bajo una simulación estocástica las conclusiones se obtienen generando repetidamente simulaciones del modelo aleatorio (Fernández Casal y Cao 2020). Una de las ventajas de la simulación estocástica es que permite experimentar una variedad de situaciones o modelos con determinadas condiciones sobre las variables que nos permite tomar una decisión al momento que se nos presenten datos reales generados por un sistema similar.

2. Metodología

2.1. Simulación de los datos

Se definieron 4 escenarios para la simulación de datos con las siguientes características:
Escenario 1: Variable respuesta altamente correlacionada con las predictoras y las variables predictoras poco correlacionadas entre sí.
Escenario 2: Variable respuesta poco correlacionada con las predictoras y las variables predictoras muy correlacionadas entre sí.
Escenario 3: Variable respuesta muy correlacionada con las predictoras y las variables predictoras también muy correlacionadas entre sí.

Escenario 4: Variable respuesta poco correlacionada con las predictoras y asimismo las variables predictoras poco correlacionadas entre sí.

En los escenarios 1 y 3, la situación planteada se corresponde con grupos "separables" en función de los valores de los predictores; mientras que en los escenarios 2 y 4 los grupos están solapados respecto a las variables predictoras, dificultando la tarea de discriminarlos en función de las mismas. Se generaron mediante simulación 500 archivos de datos para cada uno de los siguientes tamaños de muestra: 30, 75, 200, 400, 600, 1000. Se "marcó" el 20% de las observaciones para ser utilizadas como grupo de test y el restante 80% para la estimación de los modelos evaluados en cada caso. La simulación se realizó a partir de distribuciones normales multivariadas con matriz de correlaciones según cuatro estructuras diferentes (escenarios). Se consideró la primer columna (X1) como la variable respuesta dicotomizada según la mediana de la distribución y las restantes variables (X2 a X6) como las variables predictoras o explicativas. Sobre las bases simuladas se comparan las dos técnicas multivariadas de clasificación. El proceso de simulación de ficheros de datos como las aplicaciones estadísticas subsiguientes se lleva a cabo en el software R version 3.4.0 (R Core Team, 2018).

2.2. Técnicas de clasificación

Uno de los problemas de los que se ocupan las técnicas estadísticas multivariadas es la clasificación de objetos o unidades en grupos o poblaciones. Dos enfoques agrupan a las técnicas de clasificación. Uno de ellos es cuando se conocen los grupos o categorías y se pretende ubicar los individuos dentro de estas categorías a partir de los valores de ciertas variables. El segundo enfoque, que no es el utilizado en este trabajo, ocurre cuando no se conocen los grupos de antemano y se pretende establecerlos a partir de los datos observados.

Referido al primer enfoque, una de las técnicas utilizadas es Árboles de Clasificación que crean una serie de reglas basadas en las variables predictoras que permiten asignar una nueva observación a una de las categorías o grupo. Sin embargo, dado que frecuentemente se le atribuye el inconveniente del sobreajuste, surge una estimación más robusta de la mano de la técnica de Bosques aleatorios.

2.2.1. Árboles de Clasificación

Los árboles de clasificación (AC) se emplean para asignar unidades experimentales a las clases de una variable dependiente a partir de sus mediciones en uno o más predictores. En esta aplicación, los AC se emplean para asignar unidades a cada uno de los dos grupos definidos por la variable respuesta. Es un algoritmo que genera un árbol en el cual las ramas representan las decisiones y cada una de ellas genera reglas sucesivas que permiten utilizarlo para clasificar nuevas unidades.

Es un método no-paramétrico de segmentación binaria donde el árbol es construido dividiendo repetidamente los datos. En cada división los datos son partidos en dos grupos mutuamente excluyentes. Comienza el algoritmo con un nodo inicial o raíz a partir del cual se divide en dos sub-grupos o sub-nodos, esto es, se elige una variable y se determina el punto de corte de modo que las unidades pertenecientes a cada nuevo grupo definido sean lo más homogéneas posible. Luego se aplica el mismo procedimiento de partición a cada sub-nodo por separado. En cada uno de estos nodos se vuelve a repetir el proceso de seleccionar una variable y un punto de corte para dividir la muestra en dos partes homogéneas.

Las divisiones sucesivas se determinan de modo que la heterogeneidad o impureza de los sub-nodos sea menor que la del nodo de la etapa anterior a partir del cual se forman. El objetivo es particionar la respuesta en grupos homogéneos manteniendo el árbol lo más pequeño posible.

La función de impureza es una medida que permite determinar la calidad de un nodo, esta se denota por i(t). Si bien existen varias medidas de impureza utilizadas como criterios de partición, una de las más utilizadas está relacionada con el concepto de entropía

$$i(t) = \sum_{j=1}^{K} p(j/t).\ln p(j/t)$$

donde j = 1, ..., k es el número de clases de la variable respuesta categórica y p(j|t) la probabilidad de clasificación correcta para la clase j en el nodo t. El objetivo es buscar la partición que maximiza

$$\Delta i(t) = -\sum_{j=1}^{K} p(j/t).\ln p(j/t).$$

De todos los árboles que se pueden definir es necesario elegir el óptimo. El árbol óptimo será aquel árbol de menor complejidad cuya tasa de mala clasificación (TMC) es mínima. En este trabajo se registraron cuáles variables fueron elegidas por el método de construcción del árbol final, dado que no todas fueron siempre necesarias.

El desempeño del árbol se comparó mediante el porcentaje de clasificación correcta calculado sobre un conjunto de observaciones no utilizado para la construcción del mismo (datos de prueba).

2.2.2. Bosques aleatorios

Un Bosque Aleatorio consiste en un conjunto de árboles de clasificación que se combinan para dar una predicción. En esta técnica se realiza un re muestreo de los datos para estimar un conjunto de árboles y también se muestrean cuáles de las variables de entrada o explicativas participarán en el árbol.

3. Resultados

Como resultado principal se evidencia en los gráficos 1 a 4 que, en aquellos escenarios donde es favorable la clasificación por la estructura de correlaciones de las variables que suponen una separación de los grupos (Escenarios 1 y 3), la evidencia en favor de la técnica de Bosques Aleatorios es significativa, independientemente del tamaño de muestra. Sin embargo, en los casos en que la variable respuesta no fue generada correlacionada con las explicativas, y por lo tanto los grupos no son capaces de ser discriminados por los valores de dichas variables (Escenarios 2 y 4), no hay evidencia de superioridad de la técnica de Bosques excepto en aislados casos. Este comportamiento de los Bosques Aleatorio concuerda con lo que se observa al evaluar el error medio porcentual del bosque según el número de árboles estimados (Gráficos 5 y 6). Es posible distinguir comportamientos diferentes según escenario. En los escenarios más favorables para la clasificación (Escenarios 1 y 3) el error medio porcentual disminuye claramente al incrementar el tamaño del bosque y el tamaño de muestra; mientras que en casos donde la separación de grupos no es lograda por las variables explicativas, el error porcentual

medio parece mantenerse constante sin mostrar una ventaja independientemente del tamaño del bosque y del conjunto de datos.

Gráfico 1: Comparación de técnicas en Escenario 1

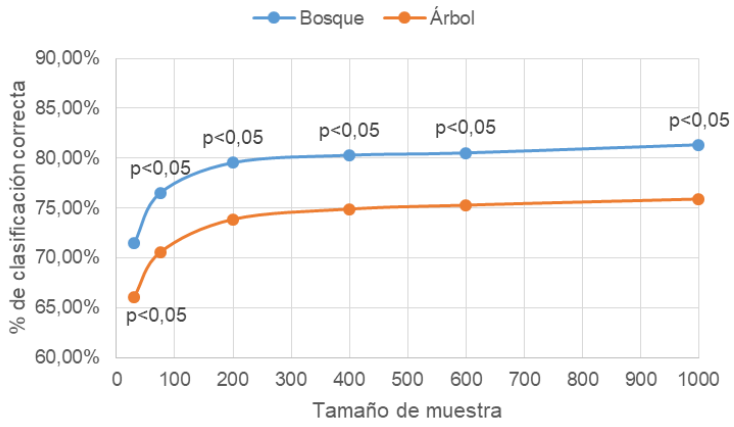


Gráfico 2: Comparación de técnicas en Escenario 2

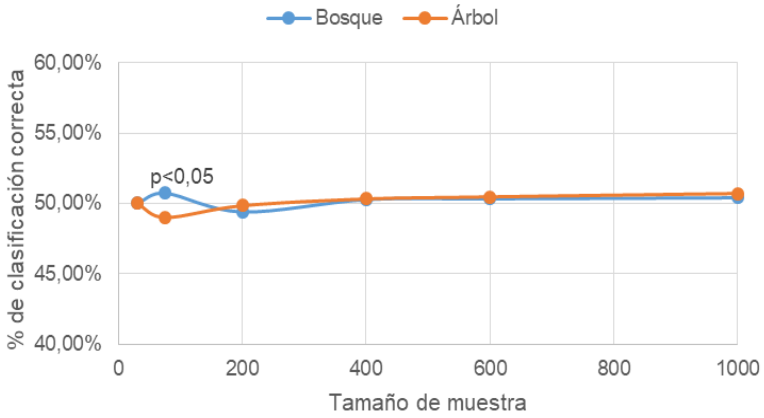


Gráfico 3: Comparación de técnicas en Escenario 3

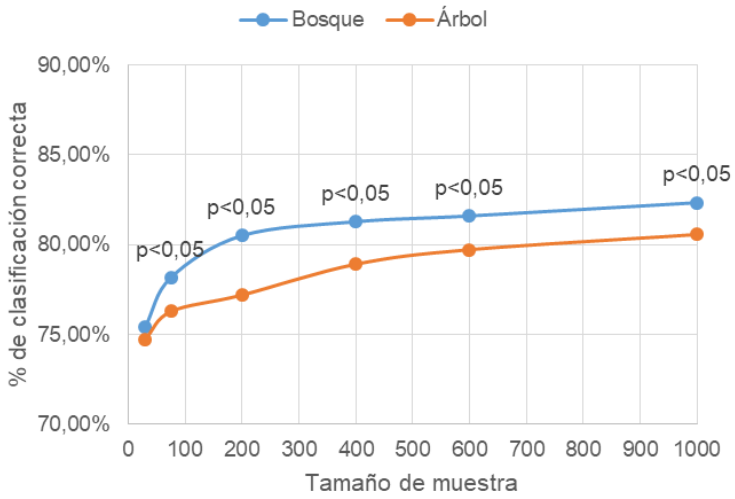


Gráfico 4: Comparación de técnicas en Escenario 4

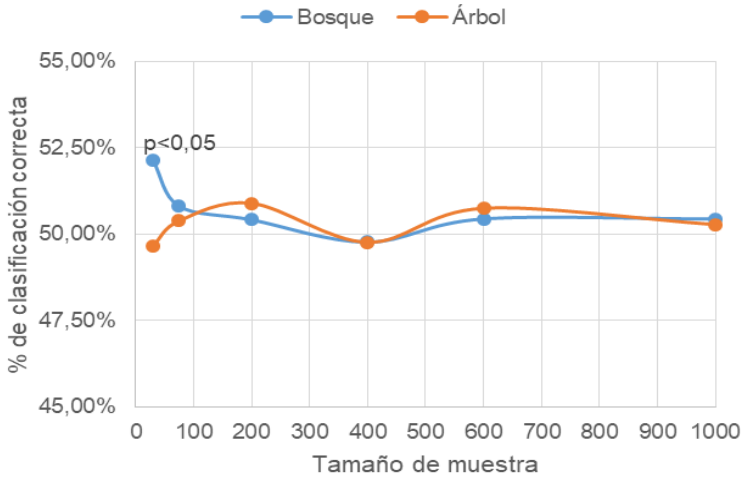


Gráfico 5: Error medio % Bosques. Escenarios 1 y 3

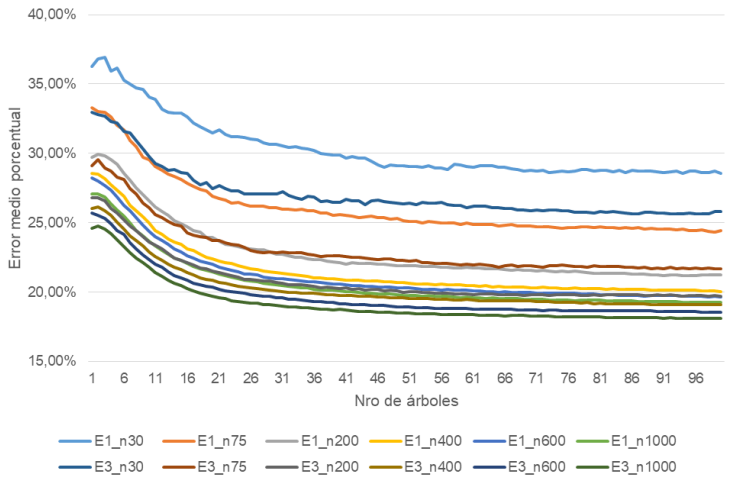
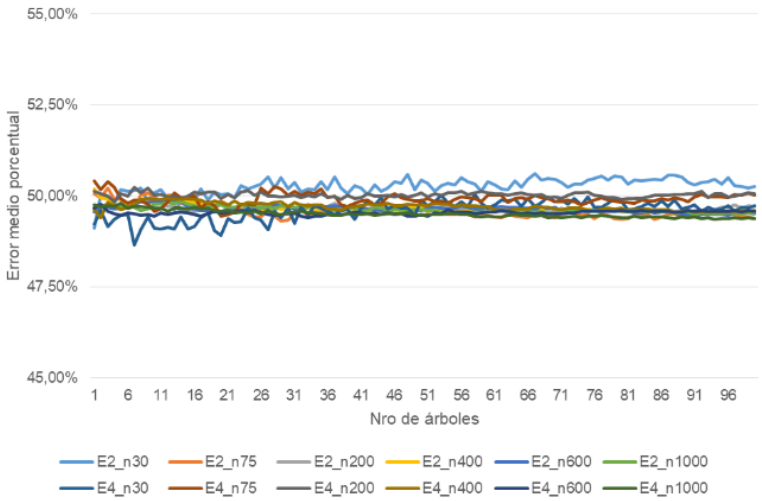


Gráfico 6: Error medio % Bosques. Escenarios 2 y 3



4. Discusión

En este trabajo se ha evaluado el desempeño de estas dos técnicas en datos simulados bajo distintas condiciones que diferían en la estructura de correlaciones entre las variables y el tamaño de muestra.

En aquellos escenarios donde es favorable la clasificación por la estructura de correlaciones de las variables que suponen una separación de los grupos, se observa un desempeño significativamente mejor para la técnica de Bosques Aleatorios, independientemente del tamaño de muestra. No obstante, en los casos en que los grupos no son capaces de ser discriminados por los valores de las variables explicativas no hay

evidencia de superioridad de la técnica de Bosques excepto en aislados casos. Similar comportamiento se observa al evaluar el error medio porcentual según el número de árboles estimados. En los escenarios más favorables para la clasificación, el error medio porcentual disminuye claramente al incrementar el tamaño del bosque y el tamaño de muestra; mientras que en casos donde la separación de grupos no es lograda por las variables explicativas, el error porcentual medio parece mantenerse constante sin mostrar una ventaja independientemente del tamaño del bosque y del conjunto de datos.

Estos resultados ponen cierta luz al momento de elegir la técnica estadística conveniente para clasificar unidades cuando las variables en consideración están o no correlacionadas y los grupos respuesta se encuentran solapados o no respecto a los valores de las mismas.

5. Bibliografía

Beltrán, C.; Barbona, I. (2020) Comparación del desempeño de técnicas multivariadas de clasificación en datos simulados bajo distintos escenarios: Regresión Logística y Árboles de Clasificación. Revista de Epistemología y Ciencias Humanas. 12 (2) 18-36

Fernández Casal, R.; Cao, R. 2020. Simulación Estadística. <https://rubenfcasal.github.io/simbook/index.html>

Hair, J.F., Anderson, R.L.,Tatham, R.L., Black, W.C. 1999. Análisis Multivariante. Prentice Hall Iberia, Madrid, España.

Hastie, T., Tibshirani, R., Friedman, J. 2009.The Elements of Statistical Learning.Data Mining, Inference, and Predicion.Springer Series in Statistics.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.