Comparación del desempeño de Árboles de clasificación y Redes Neuronales en la clasificación politómica mediante simulación.

Celina Beltrán; Ivana Barbona

Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Argentina

beltranc@dat1.net.ar

Abstract

The present work aims to study, evaluate and compare two multivariate statistical classification techniques, Neural Networks and Classification Trees, being of interest evaluate their performance when they are used in simulated data under different situations.

Data were simulated under 4 conditions that differed in the structure of correlations between the variables, each having a response variable with three categories and five continuous explanatory variables. Scenario 1 corresponds to data from a population in which the predictors are strongly correlated with the response but not with each other. Scenario 2 proposes a simulation from a population with little correlation of the response with the predictor variables but these correlated with each other. In scenario 3, the correlation present in the original population simulated is important both between the predictors and between them and the response. Finally, scenario 4 corresponds to an original population in which there is no type of correlation of significant magnitude between the variables, neither of the predictors with the response nor between them.

It was observed as the main result, that in conditions where the predictor variables are highly correlated with the response or there is multicollinearity between the predictor variables, the neural networks showed a significantly lower percentage of error in the classification. However, when the conditions for obtaining a satisfactory classification are unfavorable (predictors poorly correlated with both the response and between them), the trees achieve a percentage of correct classification that is significantly higher than the neural networks.

Keywords: neural networks; classification trees; simulation

Resumen

En esta investigación se propone el estudio, evaluación y comparación de dos técnicas estadísticas multivariadas de clasificación, Redes Neuronales y Árboles de Clasificación, siendo de interés evaluar el desempeño de las mismas cuando son utilizadas en datos simulados bajo distintas situaciones.
Se simularon datos bajo 4 condiciones diferentes que diferían en la estructura de correlaciones entre las variables, contándose en cada uno de ellos con una variable respuesta con tres categorías y cinco variables explicativas continuas. El escenario 1 corresponde a datos provenientes de una población en la que los predictores están fuertemente correlacionados con la respuesta pero no entre ellos. El escenario 2 plantea una simulación a partir de una población con poca correlación de la respuesta con las variables predictoras pero éstas correlacionadas entre sí. En el escenario 3, la correlación presente en la población origen de la simulación es importante tanto entre las predictoras como entre éstas y la respuesta. Por último, el escenario 4 corresponde a una población original en la que no existe ningún tipo de correlación de magnitud importante entre las variables, ni de los predictores con la respuesta ni entre ellos.
Se observó como resultado principal, que en condiciones donde las variables predictoras están altamente correlacionadas con la respuesta o existe multicolinealidad entre las variables predictoras, las redes neuronales mostraron un porcentaje de error significativamente menor en la clasificación. Sin embargo, cuando las condiciones para obtener una clasificación satisfactoria son desfavorables (predictores poco correlacionados tanto con la respuesta como entre ellos) los árboles logran un porcentaje de clasificación correcta notablemente superior a las redes neuronales.

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, J.F. 1999). Estas variables pueden ser cuantitativas, continuas o discretas, o cualitativas, nominales u ordinales (Pérez López, C. 2004). Uno de los objetivos de dichas técnicas es la clasificación de unidades u objetos en grupos. En la clasificación supervisada, tarea que concierne a este trabajo, 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. Las técnicas de clasificación pueden diferenciarse en aquellos métodos clásicos estadísticos y los que provienen de la Minería de datos. En las técnicas clásicas se estima un modelo estadístico cuyos coeficientes permitirán caracterizar los grupos y construir la regla de clasificación para nuevas unidades. Las inferencias sobre las estimaciones realizadas permiten detectar aquellas características que aportan en el proceso de clasificación. Esto marca una diferencia con las provenientes de la Minería de datos ya

que en estos casos generalmente los análisis son de tipo exploratorios y no se realiza una generalización sobre poblaciones de las cuales se extraen los datos.
Entre las técnicas de clasificación, correspondiente al enfoque de minería de datos respectivamente, se pueden citar: Árboles de clasificación y Redes Neuronales. En este trabajo se propone el estudio de estas dos técnicas estadísticas multivariadas de clasificación siendo de interés evaluar el desempeño de las mismas cuando son utilizadas en datos simulados bajo distintas situaciones que difieren en la estructura de correlaciones entre las variables intervinientes.

2. Metodología

2.1. Simulación de los datos

Se generaron mediante simulación 500 archivos de datos de 150 filas (unidades) y 6 columnas (variables) bajo distintas condiciones o escenarios. La simulación se realizó a partir de distribuciones normales estandarizadas multivariadas con matriz de correlaciones según cuatro estructuras diferentes. Se consideró la primer columna (X1) como la variable respuesta y las restantes variables (X2 a X6) como las variables predictoras o explicativas. Luego de la generación de los ficheros se transformaron algunas variables para crear variables categóricas con tres categorías usando los percentiles teóricos del 33% y 66% de las distribuciones para obtener grupos balanceados. La variable respuesta siempre se la consideró transformada a categórica ya que el objetivo de este estudio es evaluar las técnicas encargadas de clasificación de unidades.
Las condiciones o escenarios definidos por la estructura de la matriz de correlaciones son:

1- Escenario 1: Variable respuesta altamente correlacionada con las predictoras ($0.27<r<0.63$) y las variables predictoras poco correlacionadas entre sí ($r<0.06$).
2- Escenario 2: Variable respuesta poco correlacionada con las predictoras ($r<0.06$) y las variables predictoras muy correlacionadas entre sí ($0.49<r<0.84$).
3- Escenario 3: Variable respuesta muy correlacionada con las predictoras ($0.36<r<0.83$) y las variables predictoras también muy correlacionadas entre sí ($0.43<r<0.87$).
4- Escenario 4: Variable respuesta poco correlacionada con las predictoras y asimismo las variables predictoras poco correlacionadas entre sí ($r<0.06$).

De esta manera quedaron definidas bases de datos simulados con las siguientes características de las variables y escenarios.

- Base_3 (ESCENARIO 1): Respuesta politómica (X1) y 5 variables continuas (X2, X3, X4, X5, X6). Variable respuesta altamente correlacionada con las predictoras ($0.27<r<0.63$) y las variables predictoras poco correlacionadas entre sí ($r<0.06$).
- Base_7 (ESCENARIO 2): Respuesta politómica (X1) y 5 variables continuas (X2, X3, X4, X5, X6). Variable respuesta poco correlacionada con las predictoras ($r<0.06$) y las variables predictoras muy correlacionadas entre sí ($0.49<r<0.84$).
- Base_11 (ESCENARIO 3): Respuesta politómica (X1) y 5 variables continuas (X2, X3, X4, X5, X6). Variable respuesta muy correlacionada con las predictoras

(0.36<r<0.83) y las variables predictoras también muy correlacionadas entre sí (0.43<r<0.87).

- Base_15 (ESCENARIO 4): Respuesta politómica (X1) y 5 variables continuas (X2, X3, X4, X5, X6). Variable respuesta poco correlacionada con las predictoras y asimismo las variables predictoras poco correlacionadas entre sí (r<0.06).

Sobre las bases simuladas detalladas recientemente se comparan dos técnicas multivariadas de clasificación: ÁRBOLES DE CLASIFICACIÓN Y REDES NEURONALES. Por este motivo, para cada muestra, se simularon datos extras o suplementarios (30 filas para cada muestra) para ser considerados en la evaluación de la clasificación sin haberlos utilizados en los procesos de estimación (grupo de prueba).

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.

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 Redes Neuronales. Las Redes Neuronales son sistemas pertenecientes a una rama de la inteligencia artificial que emulan al cerebro humano y requieren un entrenamiento en base a un conocimiento previo del entorno del problema. Otra técnica aplicada frecuentemente son los 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.

2.2.1. Redes Neuronales Artificiales: El Perceptrón Multicapa

Las redes neuronales son sistemas pertenecientes a una rama de la inteligencia artificial que emulan al cerebro humano. Requieren un entrenamiento en base a un conocimiento previo del entorno del problema. Una red neuronal es un sistema compuesto por un gran número de elementos básicos, agrupados en capas que se encuentran totalmente interconectadas y que serán entrenadas para reaccionar de una determinada manera a los estímulos de entrada.

Las redes neuronales constituyen naturalmente una técnica de modelización multivariada, es decir, pueden hacer predicciones de dos o más variables simultáneamente. Pueden realizar predicciones tanto de variables continuas como discretas, utilizando las implementaciones apropiadas. El Perceptrón Multicapa (MLP, por sus siglas en inglés “Multi-Layer Perceptron”) tiene como objetivo la categorización o clasificación de forma supervisada. Utilizando el algoritmo de aprendizaje supervisado Backpropagation, la red

aprende la relación entre las variables explicativas y la categoría de pertenencia, con el propósito de lograr clasificar una nueva observación para la cual se cuenta con los valores de las variables explicativas pero se desconoce su categoría o grupo de pertenencia.

Un perceptrón multicapa está compuesto por una capa de entrada, una capa de salida y una o más capas ocultas; aunque se ha demostrado que para la mayoría de problemas bastará con una sola capa oculta. En este tipo de arquitectura, las conexiones entre neuronas son siempre hacia delante, es decir, las conexiones van desde las neuronas de una determinada capa hacia las neuronas de la siguiente capa; no hay conexiones laterales, ni conexiones hacia atrás. Este es, la información siempre se transmite desde la capa de entrada hacia la capa de salida. En dicho diagrama wji representa el peso de conexión entre la neurona de entrada i y la neurona oculta j, y vkj es el peso de conexión entre la neurona oculta j y la neurona de salida k.

En esta aplicación las P neuronas de la capa de entrada corresponden a los valores de las variables explicativas consideradas y la capa de salida estará constituida por las 3 categorías que corresponden a la variable respuesta.

Durante el aprendizaje o entrenamiento del sistema se evalúan las condiciones de pertenencia a cada una de las categorías. El aprendizaje supervisado se caracteriza por conocer la respuesta que debería tener la red frente a una determinada entrada. De esta manera, se compara la salida deseada con la salida de la red y si existen discrepancias se ajusta iterativamente los pesos considerando en cada paso la información sobre el error cometido.

El algoritmo backpropagation se basa en el ajuste de los pesos de las conexiones de la red en función de las diferencias entre los valores deseados (verdaderos) y los obtenidos por el sistema.

Así, la etapa de aprendizaje tiene por objeto hacer mínimo el error entre la salida brindada por la red y la salida deseada o verdadera. El aprendizaje se hace sobre un conjunto de datos, llamado conjunto de entrenamiento, que consta de un grupo de patrones asociados a sus correspondientes salidas.

Se pretende minimizar una función de error cuya expresión para el patrón j viene dada por

$$E_i = \frac{1}{2}\sum_{k=1}^{M}(d_{ik} - y_{ik})^2$$

donde la d_{ik} es la salida deseada para la neurona de salida k cuando se presenta el patrón i. La medida de error general se expresa como

$$E = \sum_{i=1}^{N} E_i$$

Este algoritmo realiza la modificación de los pesos basándose en la técnica del gradiente decreciente. Considerando al conjunto de pesos en un espacio de tantas dimensiones como pesos se tenga, el algoritmo busca obtener información sobre la pendiente de la superficie y modificar iterativamente los pesos de modo de hallar el mínimo global.

Una vez que se tiene la red estimada, al presentarse un patrón de entrada Xi, se transmite mediante los pesos wik desde la capa de entrada hacia la capa oculta de la red. Las neuronas de esta capa oculta aplican la función de activación a las señales recibidas obteniendo un valor de salida. Estos valores son transmitidos por los pesos vjk, quienes, mediante la aplicación de la misma función anterior, obtienen los valores de salida de la red correspondientes a las neuronas de la última capa.

Esta función de activación que se aplica sobre la entrada de cada neurona para obtener el valor de salida debe ser una función continua y derivable. En este trabajo la función de activación utilizada es del tipo sigmoidal logística.

Para realizar la validación del modelo obtenido con los datos del conjunto de entrenamiento, es necesario considerar el error que se comete cuando la red es aplicada sobre un nuevo conjunto de datos, el conjunto de prueba. Esta nueva aplicación brindará como resultado de clasificación la matriz de confusión. La matriz de confusión que muestra el tipo de las predicciones correctas e incorrectas cuando se aplica el modelo sobre el conjunto de prueba. La misma permite comprender en qué sentido se equivoca la red al intentar clasificar las nuevas observaciones. En el gráfico de esta matriz, las predicciones correctas están representadas por las barras que aparecen sobre la diagonal, mientras que el resto de las barras indican el tipo de error cometido (qué valor ha predicho el modelo y cuales el valor verdadero). La altura de las barras es proporcional al porcentaje de los registros que representan.

El desempeño del modelo se valoró mediante el porcentaje de clasificación correcta calculado sobre un conjunto de datos (datos de prueba) no utilizado para la estimación del mismo.

2.2.2. 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).

3. Resultados

3.1. Datos simulados

Cada una de las bases de datos detalladas a continuación contiene 500 muestras de 150 filas cada una:

- Base_3: Respuesta politómica (X1) y 5 variables continuas (X2, X3, X4, X5, X6). Variable respuesta altamente correlacionada con las predictoras ($0.27<r<0.63$) y las variables predictoras poco correlacionadas entre sí ($r<0.06$). Por cómo se construyó durante la simulación X1, los grupos definidos por la respuesta están aproximadamente balanceados.

- Base_7: Respuesta politómica (X1) y 5 variables continuas (X2, X3, X4, X5, X6). Variable respuesta poco correlacionada con las predictoras ($r<0.06$) y las variables predictoras muy correlacionadas entre sí ($0.49<r<0.84$). Por cómo se construyó durante la simulación X1, los grupos definidos por la respuesta están aproximadamente balanceados.

- Base_11: Respuesta politómica (X1) y 5 variables continuas (X2, X3, X4, X5, X6). Variable respuesta muy correlacionada con las predictoras ($0.36<r<0.83$) y las variables predictoras también muy correlacionadas entre sí ($0.43<r<0.87$). Por cómo se construyó durante la simulación X1, los grupos definidos por la respuesta están aproximadamente balanceados.

- Base_15: Respuesta politómica (X1) y 5 variables continuas (X2, X3, X4, X5, X6). Variable respuesta poco correlacionada con las predictoras y asimismo las variables predictoras poco correlacionadas entre sí ($r<0.06$). Por cómo se construyó durante la simulación X1, los grupos definidos por la respuesta están aproximadamente balanceados.

3.2. Aplicación de técnicas de clasificación y comparación de los resultados

Se ajustó un modelo de regresión logística y el perceptrón multicapa para variable respuesta y 5 variables explicativas continuas, para cada una de las 500 muestras en cada uno de los 4 escenarios evaluados. En cada caso se calculó el porcentaje de clasificación incorrecta y fueron comparados, respecto a las técnicas y en cada escenario, mediante el test de Wilcoxon. Los resultados hallados en cada caso se presentan a continuación.

Tabla 1: Porcentaje promedio de clasificación incorrecta según escenario y técnica estadística.

%	Árboles de clasificación					Redes Neuronales					p-valor
Escenario	Cuartil 1	Mediana	Cuartil 3	Media	DE	Cuartil 1	Mediana	Cuartil 3	Media	DE	
I	21,33	23,33	24,70	21,51	6,62	16,67	19,33	22,70	18,74	6,27	<0,0001
II	32,00	35,33	37,33	31,22	11,32	42,00	44,67	48,00	41,07	13,09	<0,0001
III	17,33	19,33	22,00	18,30	6,37	14,67	16,67	19,33	15,87	6,05	<0,0001
IV	32,00	35,33	36,67	31,90	10,08	41,33	44,00	46,00	40,80	11,51	<0,0001

En la tabla 1 se observa que el desempeño de Redes Neuronales en mejor que árboles de clasificación cuando la correlación entre la variable respuesta y las predictoras es alta,

mientras que cuando esta correlación es baja la técnica de Árboles clasifica mejor que Redes Neuronales.

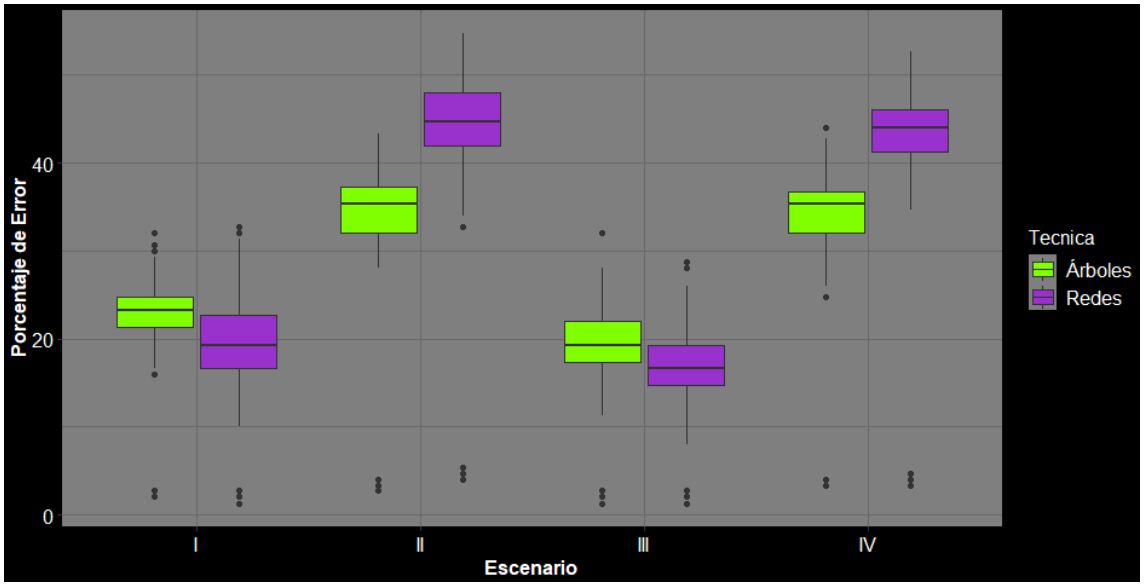

Gráfico 1: Diagrama de caja para los porcentajes medios de errores de clasificación según escenario y técnica estadística.

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 la variable respuesta y las predictoras y entre las predictoras mismas.

En condiciones donde las variables predictoras están altamente correlacionadas con la respuesta (Escenario I y III), las redes neuronales mostraron un porcentaje de error promedio significativamente menor en la clasificación, es decir, tienen un mejor desempeño que la técnica de árboles de clasificación.

Sin embargo, cuando los predictores están poco correlacionados con la respuesta (Escenario II y IV) los árboles logran un porcentaje de clasificación correcta notablemente superior a las redes neuronales, siendo su desempeño mejor que estás últimas.

En los casos en los que las variables predictoras estaban correlacionadas, es decir, cuando existe multicolinealidad (Escenario II y III) no se observa que esta situación afecte de alguna manera el desempeño de ambas técnicas.

5. Bibliografía

Agresti, A. 2002. Categorical Data Analysis. Wiley & Sons. New Jersey.

Beltrán, C. 2012 Aplicación de redes neuronales artificiales en la clasificación de textos académicos según disciplina: Biometría, Filosofía y Lingüística informática. Revista de Epistemología y Ciencias Humanas. Grupo IANUS. Rosario

Cherkassky, V., Mulier, F. 2007.Learning From Data.Concepts, Theory, and Methods.John Wiley & Sons.

Cuadras, C.M. 2014 Nuevos métodos de análisis multivariante. CMC Editions. Barcelona, España.

Hosmer, D.; Lemeshow, S. 1989. Applied Logistic Regression. Jhon Wiley & Sons. New York.

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

Witten, I., Frank, E. 2005. Data Mining.Practical Machine Learning Tools and Techniques.Elsevier.