Paquete gmtree: Predicción de estancia hospitalaria *

Juan Camilo España Lopera *Universidad de Antioquia*
Olga Cecilia Usuga Manco *Universidad de Antioquia*
Juan Sebastian Jaén Posada *Universidad de Antioquia*

Resumen: El objetivo de este tratajo es presentar la metodología de Árboles de modelos aditivos generalizados de localización, forma y escala útil en la predicción de variables cuantitativas. La metodología se implementó en el paquete gmtree del lenguaje de programación R y esta enmarcada en la predicción de estacia hospitalaria. Se destaca la creación de una estructura de modelos nuevos que se basa en un árbol de decisión y un modelo aditivo eneralizado para localización, forma y escala en cada nodo, que permite modelar fenómenos en los que se puede encontrar subgrupos de observaciones provenientes de distribuciones diferentes.

Keywords: Estancia hospitalaria, gamlss, gmtree, prediccion.

Introducción

La eficiencia en el uso de los recursos es uno de los objetivos más importantes en la literatura de administración hospitalaria. Para alcanzar este objetivo los hospitales programan adecuadamente las cirugías y otros servicios prorrogables que generarán hospitalización, de tal manera, que los recursos no estén subutilizados pero que tampoco sean insuficientes para atender la demanda. Este problema ha sido abordado por diferentes estudios por medio de la predicción de la estancia hospitalaria (Length of stay - LOS), la cual se define como la diferencia entre la fecha de egreso y la fecha de ingreso al hospital de un episodio individual de hospitalización.

El problema de predecir el LOS no ha sido investigado únicamente por su impacto en la administración hospitalaria, sino también por el interés que despierta su modelamiento y predicción en el contexto académico, debido a las características y la complejidad que estas conllevan (Turgeman, May and Sciulli, 2017).

En este trabajo se desarrolla e implementa una metodología de predicción enmarcada en la predicción de la estancia hospitalaria, en la cual se utilizan los modelos aditivos generalizados de localización, forma y escala (GAMLSS) y se incorpora un procedimiento de selección de modelos dentro de varios candidatos que se generan.

El trabajo presentado se divide en cuato secciones. En la primera sección se presenta la metodología desarrollada, luego, en la segunda sección se describen las funciones del paquete y su instalación desde el repositorio GitHub. Finalmente, en la tercera sección se presentan el desarrollo de la metodología para un caso de estudio y por último se presentan las conclusiones.

* **Autor de contacto**: camilo1260@gmail.com .

Árboles de modelos aditivos generalizados de localización, forma y escala

Con base en los aprendizajes y oportunidades de las técnicas de predicción del LOS se desarrolló la metodología Árboles de Modelos aditivos generalizados de localización, forma y escala (AMG). Con esta metodología se aprovechan los aprendizajes y ventajas encontrados con la técnica cubista, al mezclar una técnica de árboles de decisión con una técnica de regresión en los nodos, se aprovechan las oportunidades encontradas al incluir otro tipo de regresiones con la utilización de los modelos GAMLSS (Rigby and Stasinopoulos, 2005*b*) en lugar de la regresión lineal múltiple. También se define una meto-dología que incluye procedimientos que generalmente no incluyen las técnicas como pruebas de hipótesis de distribuciones, selección de distribución, selección de variables y selección de modelo final. En la selección del modelo final se incluyen criterios basados en la bondad de ajuste con penalización por complejidad como el Criterio de Información Akaike (AIC) y el Criterio de Información Bayesiano (BIC), pero también se dejan criterios basados únicamente en el comportamiento de los residuales como el Error Medio Absoluto (MAE) y la Raíz del Error Cuadrático Medio (RMSE) (España, 2019).

Finalmente, se implementa la metodología en el lenguaje de programación R (R Core Team, 2019) a través de un paquete llamado **gmtree** con diferentes argumentos que permiten al usuario desarrollar un modelo utilizando los criterios que seleccione y permitiéndole la posibilidad de incluir o no ciertos pasos de la metodología propuesta, con esta metodología se logran cubrir los principales retos encontrados en la literatura de predicción de LOS (España, 2019).

Paquete gmtree

Para la metodología descrita se desarrolló un paquete en el lenguaje de programación R que comprende un grupo de funciones que ejecuta los pasos definidos sobre un conjunto de datos. Este paquete tiene como dependencias los paquetes **rpart** (Therneau and Atkinson, 2019), **gamlss** (Rigby and Stasinopoulos, 2005*a*), **gamlss.dist** (Stasinopoulos and Rigby, 2019), **dplyr** (Wickham et al., 2019), **goftest** (Faraway et al., 2017) y **Metrics** (Hamner and Frasco, 2018) y puede consultarse en el repositorio de **GitHub** https://github.com/juancamiloespana/gmtree. El objetivo de este paquete es generar múltiples modelos AMG y seleccionar el de mejor desempeño de acuerdo al indicador definido por el usuario (AIC, BIC, MAE, RMSE). La función **gamlss_tree** ajusta modelos AMG compuestos de un árbol de decisión tipo CART y un modelo de regresión GAMLSS en cada hoja o nodo final creados por el árbol de decisión. La función **pred_gamlss** se utiliza para realizar predicciones de observaciones nuevas con base en el modelo ajustado en la función **gamlss_tree**. La función **test_dist** se utiliza para analizar el ajuste de las distribuciones del paquete gamlss que defina el usuario, utilizando dos pruebas de bondad de ajuste: Anderson Darling y Kolmogorov-Smirnov. La función **split_sample** se utiliza para dividir un conjunto de datos en dos subconjuntos, un conjunto para entrenamiento de modelos y otro para evaluación. El usuario define el porcentaje de datos que va a tener el conjunto de entrenamiento, y el porcentaje de datos restantes corresponderá al grupo de evaluación. Para la instalación del paquete **gmtree** se usan las siguientes instrucciones:

```
if (!require('devtools')) install.packages('devtools')
devtools::install_github('juancamiloespana/gmtree', force=TRUE)
```

La función principal del paquete es **gamlss_tree** y tiene los siguientes argumentos:

```
gamlss_tree(form, datos, n_dist_mod=4, var_sel="aicmodelo",
 steps=2, porc_entre=0.8, committess=1,
nom_dist=c("exGAUS","GIG","GG","BCCGo","BCPEo","G
A","GB2","BCTo","WEI3", "LOGNO", "EXP",
"PARETO2","IG","IGAMMA","NO"), cyc=50,
prueba_hip=TRUE,
acepta_h=FALSE,type="counts",
arbol_activo=TRUE)
```

Resultados

Para ejemplificar el uso de las funciones del paquete se usan los datos de **azpro** del paquete **COUNT** (Hilbe, 2016). Este conjunto de datos proviene de los archivos de pacientes cardiovasculares de Arizona de 1991. Se seleccionó un subconjunto de los campos para modelar la estancia hospitalaria de los pacientes que ingresan al hospital para recibir uno de los dos procedimientos cardiovasculares estándar: CABG y PTCA. Además del tipo de procedimiento se consideraron las covariables sexo, tipo de admisión, edad y hospital. A continuación se presentan las funciones del paquete que ejemplifican la selección de datos de entrenamiento y validación, entrenamiento del modelo y predicción del modelo.

```
# Se isntala la libreria gmtree
if (!require('devtools')) install.packages('devtools')
devtools::install_github('juancamiloespana/gmtree', force=TRUE)

# Se carga la librería COUNT para utilizar el conjunto de datos azpro
library(gmtree)
library(COUNT)
data(azpro)

#Se dividen los datos en entrenamiento y validación
sep<-split_sample(datos=azpro,perc=0.8)
train.azpro <- sep$train
test.azpro  <- sep$test
#Se usa la metodología AMG
model_amg<-gamlss_tree(los~.,datos=train.azpro)
#Se predicen los valores de LOS
pred<-pred_gamlss_tree(objeto=model_amg,newdata=test.azpro)
plot(test.azpro$los, pred$los)
```

Conclusiones

Con la presente investigación se aborda la prediccion de la estancia hospitalaria generando una metodología que utiliza métodos computacionales pero que permite el análisis de ciertas características importantes en los modelos de regresión como el ajuste de los datos a la distribución sobre la que se basa el modelo, generando dentro de la metodología AMG la versatilidad para que el usuario defina la rigurosidad con que se ajustará el modelo y sea consciente del incumplimiento de algunos supuestos para las conclusiones que se generen de estos.

References

España, J.C. 2019. Árboles de Modelos GAMLSS para la predicción de tiempo de estancia hospitalaria. Master's thesis Universidad de Antioquia Medellín, Colombia: .

Faraway, J., G. Marsaglia, J. Marsaglia and A. Baddeley. 2017. *goftest: Classical Goodness-of-Fit Tests for Univariate Distributions*. R package version 1.1-1.
URL: *https://CRAN.R-project.org/package=goftest*

Hamner, B. and M. Frasco. 2018. *Metrics: Evaluation Metrics for Machine Learning*. R package version 0.1.4.
URL: *https://CRAN.R-project.org/package=Metrics*

Hilbe, J. 2016. *COUNT: Functions, Data and Code for Count Data*. R package version 1.3.4.
URL: *https://CRAN.R-project.org/package=COUNT*

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
URL: *https://www.R-project.org/*

Rigby, R. A. and D. M. Stasinopoulos. 2005*a*. "Generalized additive models for location, scale and shape,(with discussion)." *Applied Statistics* 54:507–554.

Rigby, R. and M. Stasinopoulos. 2005*b*. "Generalized additive models for location, scale and shape." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54(3):507–554.

Stasinopoulos, M. and R. Rigby. 2019. *gamlss.dist: Distributions for Generalized Additive Models for Location Scale and Shape*. R package version 5.1-4.
URL: *https://CRAN.R-project.org/package=gamlss.dist*

Therneau, T. and B. Atkinson. 2019. *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15.
URL: *https://CRAN.R-project.org/package=rpart*

Turgeman, L., J. May and R. Sciulli. 2017. "Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission." *Expert Systems with Applications* 78:376–385.

Wickham, H., R. François, L. Henry and K. Müller. 2019. *dplyr: A Grammar of Data Manipulation*. R package version 0.8.3.
URL: *https://CRAN.R-project.org/package=dplyr*