Trabajo Práctico 1

Jairo Jiménez
Sergio De Raco

June 12, 2015

Objetivo

El objetivo del presente trabajo es analizar las característcas en las cuáles el algoritmo J48 es robusto. El algoritmo será aplicado sobre la base de datos Latinobarómetro [Lagos, 2012], la cual es una encuesta de percepción política, democrática y social de latinoamérica. Para poder observar las cualidades y deficiencias del algoritmo, este será sometido a diferentes tipos de pruebas para analizar su comportamiento.

Alcance

Resultados esperados

- Confianza: Crecimiento en training, crecimiento en test hasta un valor de confianza particular, luego decrecimiento (Empieza el sobreajuste)
- Faltantes: Robustez cuando el árbol tiene pocos faltantes o faltantes sobre una sola variable. Decadencia cuando se aumenta la cantidad de faltantes sobre todas las variables, sobre ajuste cuando la imputación es hecha teniendo en cuenta la clase, en el caso contrario se espera poco poder predictivo.
- Ruido: Poco poder predictivo del árbol en la medida que la cantidad de Ruido aumenta.
- Discretización: Robustez cuando la variable numérica no es buen discriminador. Sobre ajuste cuando la variablee numérica adquiere poder predictivo.

Metodología

Para la realización del presente trabajo, se pretende usar el sofware de minería de datos Weka [Witten and Frank, 2005] mediante el lenguaje estadístico R [R Core Team, 2013] usando el paquete RWeka [Hornik et al., 2009]

1 Sobreajuste y poda

El mejor ajuste del árbol se da cuando la confianza es de 0.1, después de este valor, empieza el sobreajuste del árbol, pues este deja baja la capacidad de pronosticar las nuevas instancias de manera correcta
A medida que aumenta la confianza, aumentan el tamaño y el número de hojas Figura 1

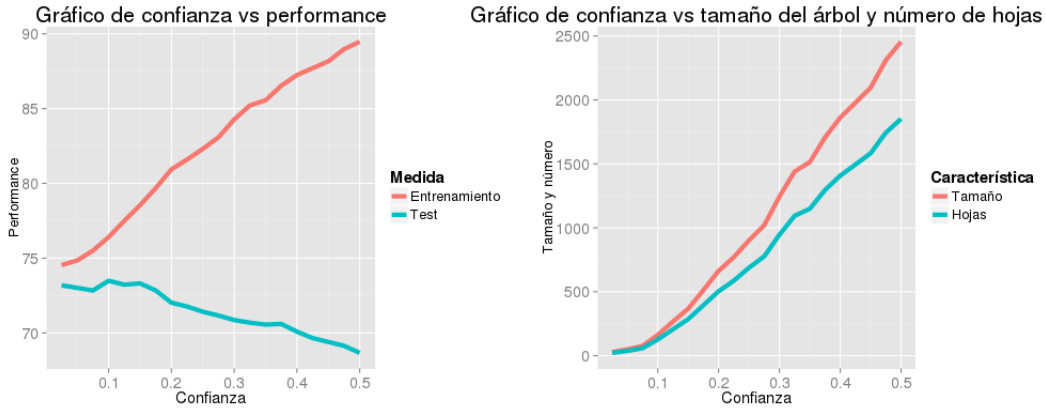

Figure 1: Gráficos de confianza

2 Tratamiento de datos faltantes

Para el análisis de datos faltantes, el algoritmo se sometió a 3 diferentes tipos de faltantes sobre el conjunto de datos. Primero se anaklizó el efecto de los datos faltantes sobre una sola columna, luego sobre varias columnas pero solamente induciendo un faltante por fila y por último sobre varias columnas pero induciendo varios faltantes por fila. En los 3 casos, la cantidad de datos faltantes fue aumentando de 0 a 85%

2.1 Sobre una sola variable

En el caso de una variable, se seleccionó como candidata la variable que separa mejor el árbol. A esta variable se le agregaron faltantes desde 0 a 85% Figura 2

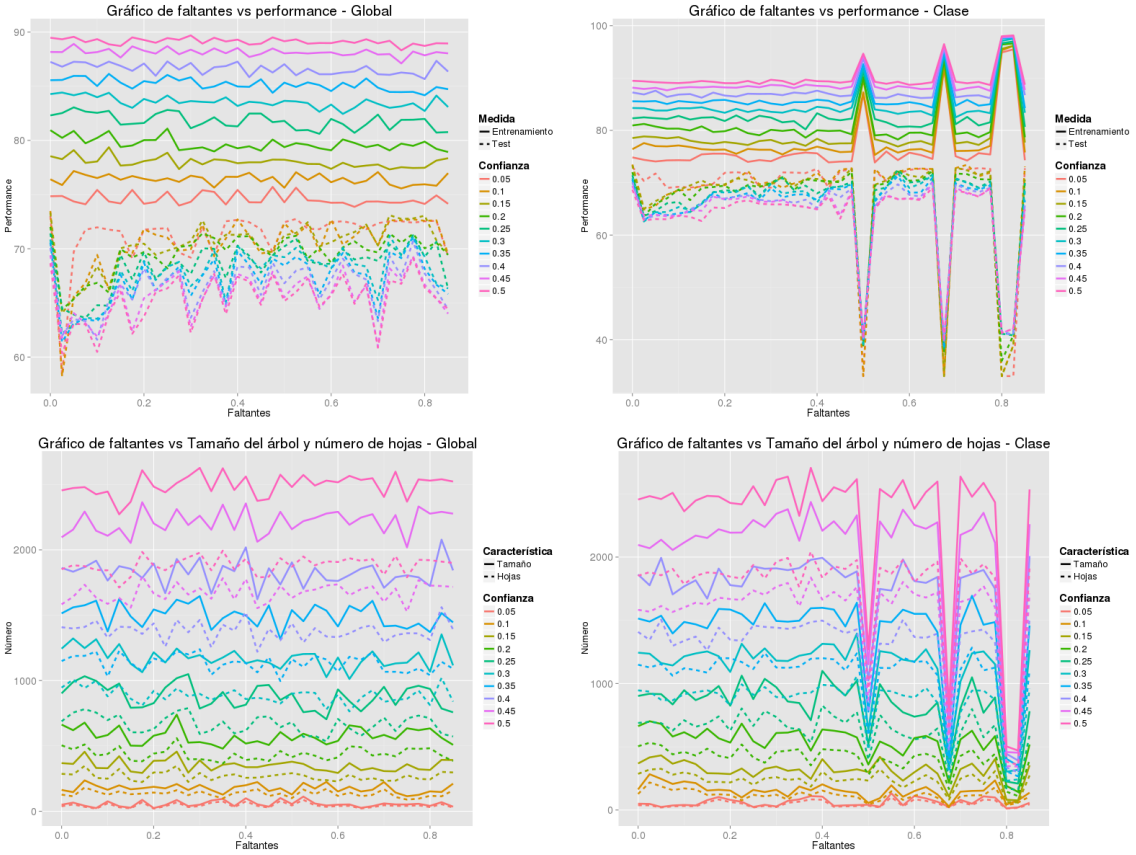

Figure 2: Gráficos de Faltantes en una sola variable

2.1.1 Imputación sin tener en cuenta la clase

2.1.2 Imputación teniendo en cuenta la clase

2.2 Sobre varias variables con un único faltante por fila

En este caso, se agregaron datos faltantes de manera aleatoria a las columnas, pero tengiendo en cuenta que solamente se admite un valor faltante por fila en el conjunto de datos. Figura 3

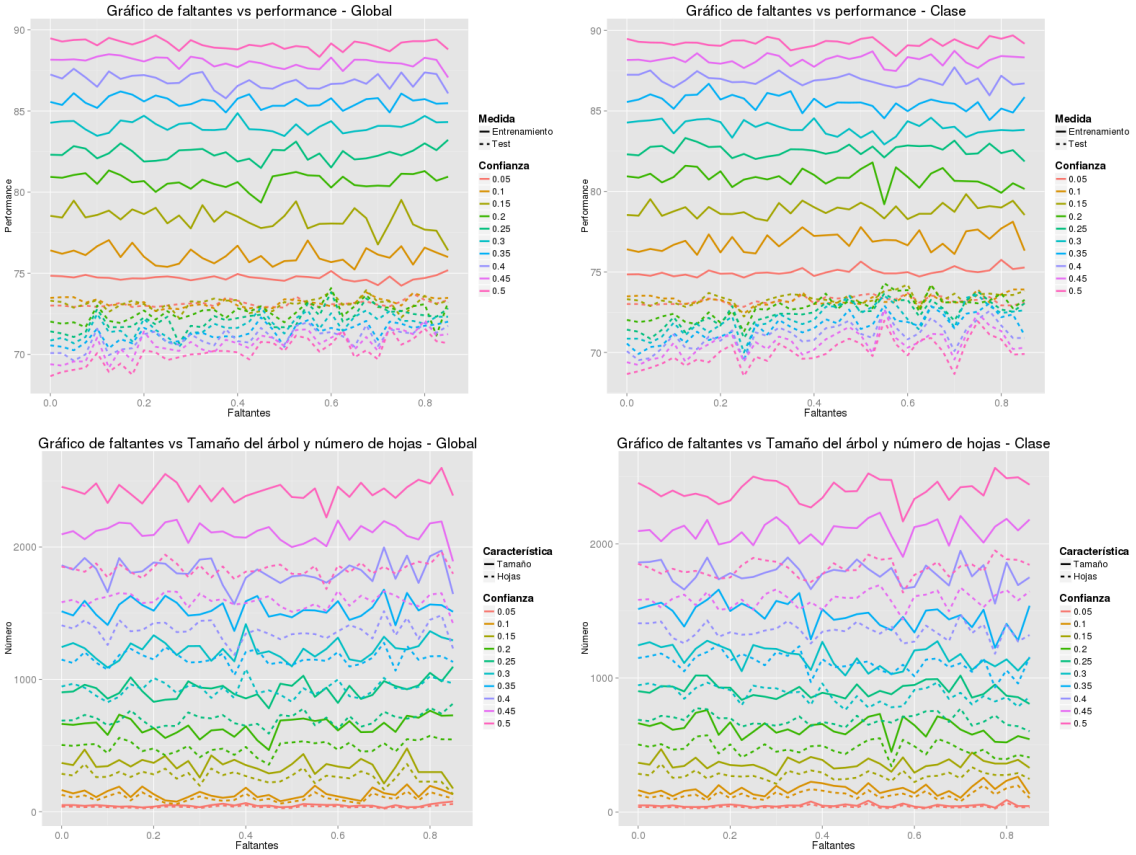

Figure 3: Gráficos de Faltantes en Múltiples variables con un solo faltante por fila

2.2.1 Imputación sin tener en cuenta la clase

2.2.2 Imputación teniendo en cuenta la clase

2.3 Sobre varias variables con múltiples faltantes por fila

Para el último análisis, se agregaron datos faltantes sobre el total de entradas de el conjunto de datos (total de filas por total de columnas). Figura 4

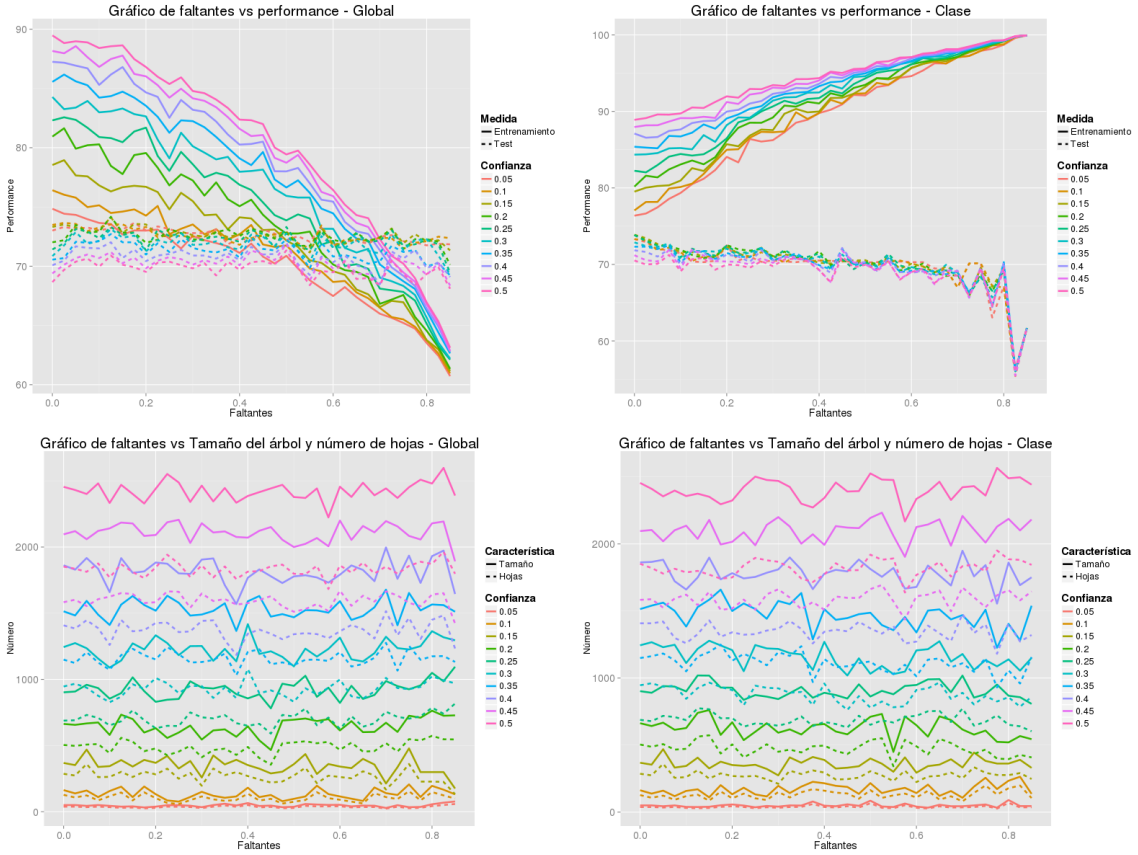

Figure 4: Gráficos de Faltantes en múltiples variables con múltiples faltantes por fila

2.3.1 Imputación sin tener en cuenta la clase

2.3.2 Imputación teniendo en cuenta la clase

3 Tolerancia al ruido

En esta parte, se genero entre 0 y 35% de ruido sobre la clase. Figura 5

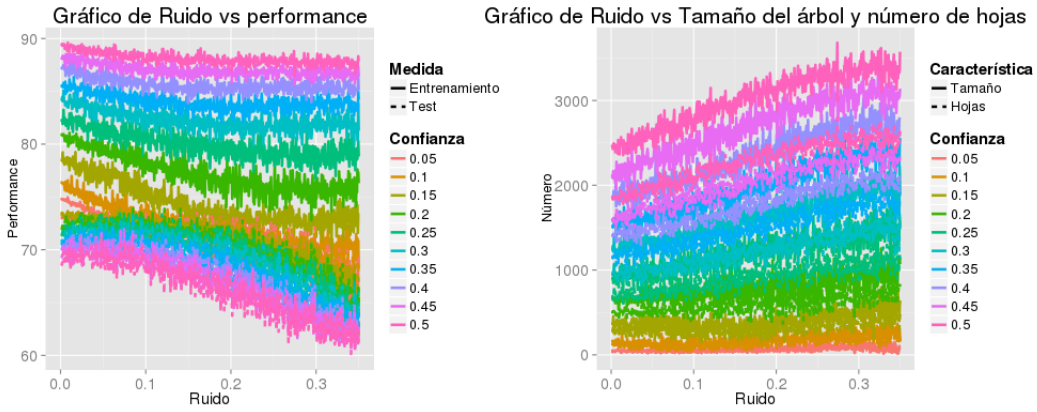

Figure 5: Gráficos de Ruido

4 Discretización de atributos numéricos

4.1 Sobre variables con poco poder discriminativo

4.2 Sobre variables con mucho poder discriminativo

Conclusiones

Con respecto a los datos faltantes, el árbol tiene una tolerancia similar a la mayoría de métodos estadísticos, en donde se admite al rededor de un 20% de datos faltantes.

Trabajo futuro

Para poder verificar las propiedades del método para su aplicación en problemas de la vida real, sería conveniente mezclar las diferentes pruebas hechas sobre en este trabajo. ¿qué pasa cuando el árbol es sometido a ruido y a faltantes al mismo tiempo?. Además de esto, sería conveniente intentar otros métodos para imputación de datos categóricos que sesguen menos los datos, como por ejemplo el método de imputación Hot Deck, el cual, busca individuos similares al individuo con el dato faltante para "donarle" el dato que le hace falta. [Rodgers, 1984]

References

[Hornik et al., 2009] Hornik, K., Buchta, C., and Zeileis, A. (2009). Open-source machine learning: R meets Weka. *Computational Statistics*, 24(2):225–232.

[Lagos, 2012] Lagos, M. (2012). Latinobarometro. opinion publica latinoamericana. Web.

[R Core Team, 2013] R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Rodgers, 1984] Rodgers, W. (1984). An evaluation of statistical matching. *Journal of Business and Economic Statistics*, page 91–102.

[Witten and Frank, 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition.