Trabajo Práctico 1

Jairo Jiménez
Sergio De Raco

June 11, 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, 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

Metodología

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

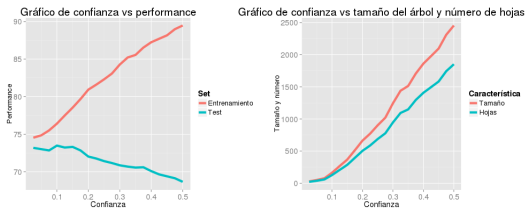

Figure 1: Gráficos de confianza

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%

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.

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).

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.