

Comparación entre J48(C4.5) y Naive Bayes

Sergio García Prado

8 de noviembre de 2016

I. INTRODUCCIÓN

La práctica consiste en la evaluación de dos algoritmos de clasificación aplicando los tests de McNemar y Student. Los algoritmos a examinar son los siguientes:

- **J48:** es un algoritmo cuya función es generar un árbol de decisión. Su nombre original es *C4.5* pero la implementación de Weka se denomina *J48*. Para generar el árbol de decisión el árbol basa la selección de atributos en cada nodo según la entropía de los mismos con respecto a la clases en la que se desea clasificar las muestras. Es por tanto un algoritmo de aprendizaje supervisado.
- **Naive Bayes:** es un algoritmo de clasificación probabilístico basado en el teorema de Bayes y algunas hipótesis adicionales que facilitan la simplificación del problema. Estas hipótesis presuponen la independencia entre variables, de ahí es de donde proviene el apelativo *naive* (*ingenuo*) ya que esta presuposición no siempre es cierta. Es un clasificador que utiliza aprendizaje supervisado y utiliza el método de máxima verosimilitud

II. TEST DE MCNEMAR: HOLDOUT DE 2/3

h_A es J48 y h_B es Naive Bayes

Número de ejemplos mal clasificados por h_A y h_B (n_{00})	Número de ejemplos mal clasificados por h_A pero no por h_B (n_{01})
Número de ejemplos mal clasificados por h_B pero no por h_A (n_{10})	Número de ejemplos bien clasificados por h_A y h_B (n_{11})

$$\frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$$

9	9
11	204

Cuadro 1: *Soybean*

4	4
13	127

Cuadro 2: *Vote*

2	2
0	16

Cuadro 3: *Labor*

III. TEST DE STUDENT: CROSS VALIDATION SIN REPETICIÓN

IV. TEST DE STUDENT: CROSS VALIDATION CON REPETICIÓN

V. RESULTADOS

Los resultados obtenidos según los test realizados con los conjuntos de datos y los tipos de test son los siguientes:

	McNemar	Student	Student (C)	Student rep.	Student rep.(C)
Soybean	J48	NB	NB	NB	NB
Vote	J48	J48	J48	J48	J48
Labor	NB	NB	NB	NB	NB