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

Denominaremos h_A al clasificador J48 y h_B a Naive Bayes.

Los conjuntos de datos que se han utilizado en esta comparación son los siguientes:

- **Soybean**: 683 muestras de 35 atributos y 19 clases.
- **Vote**: 435 muestras de 16 atributos y 2 clases.
- **Labor**: 57 muestras de 16 atributos y 2 clases.

II. Test de McNemar

El test de McNemar es un test no paramétrico, es decir, no asume ninguna distribución subyacente en el conjunto de datos. Este test se utiliza cuando tan solo se puede realizar una única ejecucción. Se basa en un test χ^2 con 1 grado de libertad y un 95 % de confianza.

Se utiliza como estadístico $\frac{(|n_{01}-n10|-1)^2}{n_{01}+n_{10}}$ siendo cada una de las variables lo indicado en la siguiente tabla:

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})

La hipótesis nula que se utiliza es que los dos clasificadores tengan la misma tasa de error,es decir $n_{01} = n_{10}$. Por tanto, para rechazar dicha hipótesis el valor calculado por el estadístico debe ser mayor que $\chi^2_{1,0,95} = 3{,}841459$. Para que los resultados de este test sean confiables $n_{01} + n_{10} > 25$. Este test tiene un bajo error Tipo 1 y un alto error Tipo 2, lo que se traduce en aceptar la hipótesis de igualdad de clasificadores cuando en realidad no lo son.

Para realizar este test se debe tener 2 particiones en el conjunto de muestras, una destinada a la fase de entrenamiento y otra a la de test. El método utilizado en esta comparación es un **HoldOut de** 2/3 utilizando las misma particiones para los dos clasificadores. Las tablas para realizar el test de McNemar con cada conjunto de datos son las siguientes siendo h_A = J48 y h_B = Naive Bayes:

9	9
11	204

Cuadro 1: *Soybean*

4	4
13	127

Cuadro 2: *Vote*

2	2
0	16

Cuadro 3: *Labor*

Como vemos, ninguno de los resultados cumple la condición de confiabilidad $n_{01}+n_{10} > 25$. A pesar de ello procederemos a realizar el test pero teniendo en dicho factor a la hora de analizar los resultados obtenidos:

- *Soybean*: $\frac{(|9-11|-1)^2}{9+11} = 0{,}05$ por lo que no rechazamos la hipótesis.
- *Vote*: $\frac{(|4-13|-1)^2}{4+13} = 3{,}76470588235$ por lo que no rechazamos la hipótesis (por poco).
- *Labor*: $\frac{(|0-2|-1)^2}{0+2} = 0{,}5$ por lo que no no rechazamos la hipótesis.

III. Test de Student

El test de Student se utiliza cuando si que es posible realizar varias ejecucciones del test. En este caso, el test si que es paramétrico. Al igual que en el caso anterior, lo que se pretende probar es que las tasas de error no son significativamente diferentes. No se conoce la varianza pero se estima que la media sigue una distribución t-student.

El estadístico utilizado en este test es $t = \frac{\overline{d}}{\sqrt{\frac{S^2_d}{kxR}}}$ donde $\overline{d} = \overline{x} - \overline{y}$. siendo x e y cada una de los conjuntos de datos, k el número de muestras y R el de repeticiones. Este test tiene un error Tipo 1 aceptable y un bajo error Tipo 2, lo que se traduce en rechazar la hipótesis de igualdad de clasificadores cuando en realidad lo son.

Para que el test sea válido los conjuntos de datos deben ser independientes entre si, pero dado que en la práctica no se dispone de ello se suele utilizar CrossValidation para tratar de obtenerlos. A este test se le puede aplicar una heurística (suma de una constante (instancias de prueba/instancias de entrenamiento)) para tratar de mejorar su precisión, a lo cual se denomina corrección.

Para la realización de esta comparación se ha utlizado **CrossValidation de 10 particiones** y se ha realizado el test de student de las siguientes formas:

- Test de Student sin repetición
- Test de Student sin repetición corregido
- Test de Student con 10 repeticiones
- Test de Student con 10 repeticiones corregido

IV. 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	–	–	–	NB	–
Vote	–	J48	J48	J48	J48
Labor	–	NB	–	NB	NB

Como se puede apreciar en la tabla de resultados, la mayoría de los test proporcionan ventaja al mismo clasificador para el mismo conjunto de datos. A pesar de ello, el conjunto de datos *Soybean* es el que más dificultades genera a los tests a la hora de detectar cuál de los clasificadores presenta mejores resultados. Probablemente la causa de ello sea debida a la cantidad de clases que tiene este conjunto de datos con respecto a los otros (19 frente a 2).

En cuanto a los conjuntos *Vote* y *Labor* se puede ver una tendendencia clara por los distintos tests a un determinado algoritmo de clasificación. Además tiene sentido la victoria de cada clasificador en cada caso, ya que Naive Bayes (por ser probabilista) se comporta mejor con atributos numéricos (*Labor*) y J48 (por ser jerárquico) para atributos binarios (*Vote*)

Las celdas que se muestran en blanco en la tabla representan los casos en los cuales la hipótesis nula no ha sido rechazada, es decir, cuando la tasa de error de ambos clasificadores no es significativamente diferente. En el caso del test de McNemar el resultado no nos proporcionó independencia en cuanto a las tasas de error, pero debemos recordar que el conjunto de datos era demasiado pequeño para poder asegurarlo con un alto grado de confianza (En el caso de *Vote* estuvo a punto de permitir que se rechazara la hipótesis nula por lo que habría dado la victoria a J48).

Por último, desde la perspectiva de rendimiento del test, el que da resultados en todos los casos es Student con Repetición, mientras que el test de McNemar no nos proporciona resultados confiables en ningún caso. Una causa de este suceso puede ser debido al tamaño de los conjuntos de datos.