

PRÁCTICA 4:

COMPARACIÓN DE DOS MÉTODOS DE APRENDIZAJE SOBRE UN CONJUNTO DE DATOS

**MINERÍA DE DATOS
PATRICIA AGUADO LABRADOR**

Conjunto de datos: Soybean
 683 instancias
 36 atributos (35 + clase)
 19 clases

TEST DE MCNEMAR

Dividimos el conjunto de datos utilizando el filtro no supervisado sobre instancias *resample*, 2/3 para entrenamiento y 1/3 para prueba, de este modo obtenemos 455 instancias y 229 respectivamente.

hB = SVM			
hA = J48			
		Mal hB	Bien hB
Mal hA		10	14
Bien hA		3	202

No es aplicable ya que $n_{01} + n_{10} = 14 + 3 = 17$, que no es mayor que 25.

VALIDACIÓN CRUZADA SIN REPETICIÓN Y TEST DE STUDENT

Tasa de error de NB = 0,0704			
	J48	IBK1	SVM(lineal)
tS	0,0849	0,0879	0,06
tS corregido	0,0849	0,0879	0,06

Al tomar NB como algoritmo base, ambos test de student muestran que no hay diferencias significativas entre los algoritmos (vemos que las tasas de error son bastante parecidas, salvo SVM que mejoran un poco).

Tasa de error de SVM(lineal) = 0,06			
	J48	IBK1	NB
tS	0,0849 (v)	0,0879 (v)	0,0704
tS corregido	0,0849 (v)	0,0879 (v)	0,0704

Al tomar SVM como algoritmo base vemos que NB es el algoritmo ganador frente a los demás. Vemos que aunque las tasas de error sean parecidas, las de Naive Bayes son las que más se acercan a la que se obtuvo con SVM lineal.

VALIDACIÓN CRUZADA CON REPETICIÓN Y TEST DE STUDENT

Tasa de error de NB = 0,0482			
	J48	IBK1	SVM(lineal)
tS	0,0561 (v)	0,0601 (v)	0,047
tS corregido	0,0561	0,0601 (v)	0,047

Al emplear validación cruzada repetida vemos que en general las tasas de error mejoran mucho, en este caso empleamos NB como algoritmo base y vemos que ambos test difieren en cuanto al algoritmo J48, solo el test corregido considera que tiene diferencias significativas con los demás. Ambos test consideran que SVM es el mejor ya que las tasas de error prácticamente son iguales a las de NB.

Tasa de error de SVM(lineal) = 0,047			
	J48	IBK1	NB
tS	0,0561 (v)	0,0601 (v)	0,0482
tS corregido	0,0561	0,0601	0,0482

En el caso de utiliza SVM como algoritmo base obtenemos con ambos test que NB es el mejor algoritmo. Además en este caso el test de Student no considera que haya diferencias significativas entre los algoritmos.

Conclusiones

Vemos que al emplear validación cruzada repetida todas las tasas de error mejoran bastante con respecto a validación cruzada sin repeticiones. El test de Student corregido es más estricto que el test sin corregir y podemos verlo ya que el segundo acepta con más facilidad que un algoritmo sea mejor que otro.

El test de McNemar no se pudo aplicar.