Práctica metodología experimental: creación y evaluación de hipótesis (y poda de árboles)

Jhon Steeven Cabanilla Alvarado

1. Descripción de los conjuntos de datos

- Soybean: formado por 683 instancias formadas por 35 atributos, todos ellos de carácter nominal. La clase de destino puede tomar 19 valores distintos. El conjunto de datos se corresponde con instancias referidas a atributos de plantas y la clase de destino representa el tipo de planta.
- Vote: formado por 435 instancias formadas por 16 atributos, todos ellos de carácter nominal. La clase de destino puede tomar 2 valores distintos. El conjunto de datos se refiere a resultados de encuestas a ciudadanos estadounidenses para tratar de predecir si votarán al partido demócrata o republicano.

2. Descripción de los algoritmos (no hay que describir C4.5, es un estándar; no hay que describir Weka, pero sí indicar las herramientas que utilizáis)

- **50T**: Para quedarme con 50 estancias uso la opción de *Weka* 'Percentage split'. Teniendo en cuenta el total de datos de cada set, para soybean usaré el 7.325% y para vote el 11.5%.

- **Hold Out**: Para este método se divide el conjunto total de datos en 2, el de prueba y el de entrenamiento. Como el enunciado dice ⅓ y ⅔ uso la opción de antes indicando un 66%.

- **Hold Out Repetido**: Cambiando la semilla de selección de los conjuntos en Weka por 2, 3 y 4 respectivamente saco más valores y con todos ellos calculo las tasas medias de error y demás.

- **Validación Cruzada**: Uso la opción de Weka 'Cross-validation' dejando 10 en el parámetro del número de particiones como indica el enunciado. En este caso cada subconjunto de particiones se usará como el de prueba y el resto como el de entrenamiento.

- **Validación Cruzada Repetida**: Igual que con el Hold Out repito el proceso del anterior cambiando las semillas de selección de subconjuntos por 2, 3 y 4.

3. Experimentos con las muestras de 50 instancias de cada conjunto de datos (ejercicio previo)

Datos	Algoritmo	Método: 50T, resto			
		Tasa error	Desviación estándar	Intervalo superior	Intervalo inferior
Soybean_50	J48	0.5024	0.0199	0.5414	0.4634
	Sin podar	0.4550	0.0198	0.4938	0.4162
Vote_50	J48	0.0468	0.0108	0.0679	0.0257
	Sin podar	0.0545	0.0116	0.0772	0.0318

4. Experimentos de hold out sin repetición

Datos	Algoritmo	Método: Hold out			
		Tasa error	Desviación estándar	Intervalo superior	Intervalo inferior
Soybean_50	J48	0.0948	0.0195	0.1330	0.0566
	Sin podar	0.1207	0.0217	0.1632	0.0782
Vote_50	J48	0.0270	0.0108	0.0482	0.0058
	Sin podar	0.0270	0.0108	0.0482	0.0058

5. Experimentos de hold out con repetición

Datos	Algoritmo	Tasa de error		
		2	3	4
Soybean_50	J48	0.1121	0.1078	0.1379
	Sin podar	0.1078	0.1078	0.1422
Vote_50	J48	0.0811	0.0541	0.0608
	Sin podar	0.0676	0.0541	0.1081

Datos	Algoritmo	Método: Hold out repetido			
		Tasa error	Desviación estándar	Intervalo superior	Intervalo inferior
Soybean_50	J48	0.1132	0.0181	0.1344	0.0919
	Sin podar	0.1196	0.0162	0.1387	0.1005
Vote_50	J48	0.0558	0.0223	0.0820	0.0295
	Sin podar	0.0642	0.0338	0.1040	0.0244

6. Experimentos de validación cruzada sin repetición

Datos	Algoritmo	Método: 10 XV			
		Tasa error	Desviación estándar	Intervalo superior	Intervalo inferior
Soybean_50	J48	0.0822	0.0107	0.0929	0.0715
	Sin podar	0.0922	0.0095	0.1017	0.0827
Vote_50	J48	0.0343	0.0057	0.0400	0.0286
	Sin podar	0.0424	0.0088	0.0512	0.0336

7. Experimentos de validación cruzada con repetición

Datos	Algoritmo	Tasa de error		
		2	3	4
Soybean_50	J48	0.0981	0.0908	0.0791
	Sin podar	0.0864	0.1010	0.0893
Vote_50	J48	0.0322	0.0368	0.0345
	Sin podar	0.0506	0.0575	0.0529

Datos	Algoritmo	Método: 10 XV			
		Tasa error	Desviación estándar	Intervalo superior	Intervalo inferior
Soybean_50	J48	0.0876	0.0086	0.0977	0.0774
	Sin podar	0.0922	0.0063	0.0996	0.0848
Vote_50	J48	0.0345	0.0019	0.0367	0.0322
	Sin podar	0.0509	0.0063	0.0583	0.0434

8. Tablas comparativas y discusión de resultados

- Soybean**:**

Algoritmo	50 instan. entrenam.	Hold out	Hold out repetido	10-XV	4 x 10 XV
J48					
Error	0.5024	0.0948	0.1132	0.0822	0.0822
Desviación	0.0199	0.0195	0.0181	0.0107	0.0081
Intervalo superior	0.5414	0.1330	0.1344	0.0929	0.0978
Intervalo inferior	0.4634	0.0566	0.0919	0.0715	0.0787
Sin podar					
Error	0.4550	0.1207	0.1196	0.0922	0.0893
Desviación	0.0198	0.0217	0.1196	0.0095	0.0086
Intervalo superior	0.4938	0.1632	0.1387	0.1017	0.0994
Intervalo inferior	0.4162	0.0782	0.1005	0.0827	0.0792

- Vote:

Algoritmo	50 instan. entrenam.	Hold out	Hold out repetido	10-XV	4 x 10 XV
J48					
Error	0.0468	0.0270	0.0558	0.0343	0.0351
Desviación	0.0108	0.0108	0.0223	0.0057	0.0022
Intervalo superior	0.0679	0.0482	0.0820	0.0400	0.0377
Intervalo inferior	0.0257	0.0058	0.0295	0.0286	0.0325
Sin podar					
Error	0.0545	0.0270	0.0642	0.0424	0.0523
Desviación	0.0116	0.0108	0.0338	0.0088	0.0039
Intervalo superior	0.0772	0.0482	0.1040	0.0512	0.0569
Intervalo inferior	0.0318	0.0058	0.0244	0.0336	0.0477

J48, 50T: tiene la mayor tasa de error de J48 entre todos los métodos. Es razonable, pues al entrenar con menos instancias, es de esperar que cree un peor modelo. La desviación no es mayor porque el conjunto de prueba tiene más instancias.
El primer experimento de hold-out estima la tasa de error más baja. Esto es engañoso, pues hold out muestra una gran variabilidad, como se aprecia con hold-out repetido, que es el que tiene mayor varianza e intervalos más amplios. Se debe a la partición aleatoria. Los procedimientos de validación cruzada estiman las menores tasas de error, como es de esperar al entrenar con más datos. Son menos variables y tienen intervalos más pequeños. Si hubiésemos repetido 10 veces la validación cruzada, el intervalo sería aún menor. La diferencia entre repetir hold out y repetir XV 4 veces también es notable. La diferencia entre y hold out y XV sin repetición también lo es, generando intervalos de confianza más pequeños, por mucho que la tasa estimada por hold out sea menor. J48 sin podar funciona mejor que J48 sobre los conjuntos más pequeños. Parece que hacen falta más instancias para que la poda sea efectiva. La diferencia es más notable con 50 instancias: no hay suficientes instancias para que sobre ajuste demasiado sin podar, y la poda empeora porque las estimaciones del error sobre los 50 datos que se usan para podar son peores. La diferencia disminuye con *hold-out* y con validación cruzada ya se aprecia que podar es ventajoso. También se aprecia que sin podar, validación cruzada no disminuye mucho la variabilidad frente a no repetir, posiblemente por su tendencia a sobre ajustar, generando clasificadores más diferentes con pequeñas variaciones del conjunto de datos.

9. Preguntas sobre validación cruzada
¿Qué tasa de error se obtendría con el método 2?
Debería ser la misma.

¿Cómo espera que varíe la estimación de la varianza con el método 2 frente al método 1?
Se espera que su valor sea inferior.

¿Y los intervalos de confianza?
Se espera que sean más pequeños.

10. Referencias
Apuntes de TAA.