Práctica 4: Metodología Experimental

Asignatura: Técnicas de aprendizaje automático (TAA)

Profesor: Teodoro Calonge

Alumno: Miguel Chaveinte García

1. Introducción

El motivo principal por el cual se realiza este conjunto de experimentos es la comparación de las distintas tasas de error mediante cada una de las técnicas, tratando de apreciar el sesgo que producen cada una de ellas, así como la variación que producen. Las técnicas que utilizadas han sido:50 instancias de entranamiento, Holdout 2 /3 , 1/3 ;3 Repeticiones de Holdout 2 /3 , 1/3 ; Validación Cruzada de 10 capas y 3 Repeticiones de Validación Cruzada de 10 capas. Dichas metodologías experimentales se describirán en cada una de sus correspondientes secciones. A continuación, se describen brevemente los algoritmos y conjuntos de datos utilizados para las labores experimentales.

2. Algoritmos

El algoritmo utilizado para las tareas de aprendizaje pertenece a la categoría de Aprendizaje Inductivo Basado en el Error. Se basa en Aprendizaje Supervisado, es decir, en la fase de entrenamiento utilizan el valor de la clase de destino como medida del error, el cual tratan de reducir al máximo. Mediante dicha estrategia tratan de conseguir clasificar correctamente las instancias futuras.

- J48: Es la implementación en Java de C4.5, un método de generación de árboles de decisión basado en la Teoría de la Información. En cada iteración trata de maximizar la ganancia de información producida tras cada partición con respecto de la clase de destino. Además, proporciona otras mejoras como poda de ramas para evitar el sobreajuste, el uso de valores continuos o el tratamiento de valores desconocidos.

Ahora voy a comentar que he realizado para cada una de las tablas:

- Método 50T, resto: Para únicamente quedarme con las 50 instancias que se pide modifico la opción de Percentaje Split, y para ello tengo en cuenta el número total de instancias que tiene cada archivo para soybean introduzco el 7,325% y para vote el 11.5%.
- Hold Out: Ahora el conjunto total de los datos se divide en 2, el conjunto de los datos que se utiliza para prueba y el que se usa para entrenamiento, por tanto, en Percentaje Split, como te pide un ⅓ y ⅔ he introducido 66%. Dejando las semillas en 1 como indica el enunciado.
- Hold Out repetido: Igual que antes, pero en el apartado de More Options modifico el valor de la semilla introduciendo 2-3-4 y tomando los datos para completar la tabla. Para la tabla de Hold Out repetido, saco la media de las tasas de error, es decir sumo la tasa de error obtenida para la semilla 1-2-3-4 y con esa nueva tasa de error hallo la desviación típica,utilizando en esta

última la cuasi-varianza, la cúal reduce el sesgo que se podría producir respecto del valor poblacional.En el caso de los intervalos de confianza hemos utilizado la fórmula que se apoya en la distribución T de Student, así como la media de la tasa de error y la desviación típica calculada anteriormente.

- Validación Cruzada: Ahora en vez de seleccionar Percentaje Split uso la opción de CrossValidation dejando 10 en el parámetro.
- Validación Cruzada Repetida: hago lo mismo que antes, pero al igual que para el método Hold Out, repito modificando en el apartado More Options la semilla introduciendo 2-3-4 y apuntando los datos en la tabla. Al igual que antes para la tabla de Validación Cruzada Repetida saco la media de las tasas de error, es decir sumo la tasa de error obtenida para la semilla 1-2-3-4 y con esa nueva tasa de error hallo los intervalos y la desviación típica.

3. Conjunto de datos

Se han utilizado 3 conjuntos de datos en los experimentos realizados. Estos se describen brevemente a continuación:

- Soybean: Está 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: Está 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. En las siguientes secciones se describen los experimentos realizados, así como los resultados obtenidos en cada caso junto con una discusión acerca de los mismos. Algo a destacar es el uso de distintas semillas para la tarea de particionamiento, las cuales se han indicado en las tablas de resultados según corresponda.

4. Fórmulas utilizadas

- Para 50T y HOLD OUT 2 /3, 1/3

$$e_S(h) = r/n$$

$$\sigma_{e_s(h)} \approx \left(\frac{e_{S(h)}(1-e_{S(h)})}{n}\right)^{1/2}$$

$$e_{S(h)} \pm z_N \times \left(\frac{e_S(h)(1-e_S(h))}{n}\right)^{1/2}$$

- Para Hold Out (Rep.) o Validación Cruzada:

$$e(h) = \sum_{i=1,k} e_i(h)/k$$

$$S^2_{e(h)} = 1/(k-1) \times \sum_{i=1,k} (e_i(h) - e(h))^2$$

$$e(h) \pm t_{N,k-1} \times S_{e(h)}/\sqrt{k}$$

- Para Validación Cruzada (Rep.)

$$e(h)=[\sum_{i=1,R\times k} e_i(h)]/(R \times k)$$

$$S^2_{e(h)} = 1/(R\times k-1) \times \sum_{i=1,R\times k} (e_i(h)- e(h))^2$$

$$e(h) \pm t_{N,R\times k-1} \times S_{e(h)}/\sqrt{R\times k}$$

5. Ejercicio inicial 50T

Datos	Algoritmo	Método: 50T, resto (semilla 10)			
		Tasa error	Desviación estandar	Intervalos	
Soybean_50	J48	0,459716	0,01980861	0,420891124	0,498540876
	Sin podar	0,46445	0,019822921	0,425597076	0,503302924
Vote_50	J48	0,41558	0,025116521	0,366351618	0,464808382
	Sin podar	0,05974	0,012078868	0,036065419	0,083414581

6. Hold out 2/3,1/3

Datos	Algoritmo	Método: hold out (semilla 1)			
		Tasa error	Desviación estandar	Intervalos	
Soybean_50	J48	0,094828	0,019417089	0,056770505	0,132885495
	Sin podar	0,133621	0,022549743	0,089423504	0,177818496
Vote_50	J48	0,027027	0,013466828	0,000632017	0,053421983
	Sin podar	0,027027	0,013466828	0,000632017	0,053421983

7. Hold out 2/3,1/3 (repetido)

Datos	Algoritmo	Hold out diferentes semillas		
		Tasa error semilla 2	Tasa error semilla 3	Tasa error semilla 4
Soybean_50	J48	0,112069	0,107759	0,137931
	Sin podar	0,116379	0,12931	0,142241
Vote_50	J48	0,081081	0,054054	0,060811
	Sin podar	0,067568	0,054054	0,060811

Datos	Algoritmo	Método: Hold out repetido			
		Tasa error	Desviación estandar	Intervalos	
Soybean_50	J48	0,11314675	0,018074153	0,102512822	0,123780678
	Sin podar	0,13038775	0,0107759	0,124047749	0,136727751
Vote_50	J48	0,05574325	0,022324577	0,042608585	0,068877915
	Sin podar	0,052365	0,017770135	0,041909941	0,062820059

8. Validación cruzada 10 particiones

Datos	Algoritmo	Método: 10 XV			
		Tasa error	Desviación estandar	Intervalos	
Soybean_50	J48	0,0849	0,024731342	0,036437056	0,133383916
	Sin podar	0,0864	0,027349989	0,032839035	0,140050991
Vote_50	J48	0,036680761	0,034235439	-0,0304207	0,103782222
	Sin podar	0,036786469	0,034424408	-0,03068537	0,104258308

9. Validación cruzada 10 particiones (repetido)

Datos	Algoritmo	Validacion cruzada repetida semillas		
		Tasa error semilla 2	Tasa error semilla 3	Tasa error semilla 4
Soybean_50	J48	0,0981	0,0908	0,0791
	Sin podar	0,1054	0,1026	0,0893
Vote_50	J48	0,032241015	0,036997886	0,034513742
	Sin podar	0,043657505	0,041596195	0,039164905

Datos		Algoritmo	Método: Validación cruzada repetida			
			Tasa error	Desviación estandar	Intervalos	
Soybean_50		J48	0,088235	0,00812996	0,083452032	0,093018556
		Sin podar	0,0959	0,009442595	0,090379019	0,10149012
Vote_50		J48	0,035108351	0,002207411	0,033809621	0,036407081
		Sin podar	0,040301268	0,002976928	0,038549793	0,042052744

10. Tablas Comparativas

Soybean_50										
Algoritmo	50 instancias entrenamiento		Hold out		Hold out repetido (4)		10-XV		4x 10-XV	
J48										
Error	0,459716		0,094828		0,11314675		0,0849		0,0882	
Desviación	0,01980861		0,019417089		0,018074153		0,024731342		0,00812996	
Intervalos	0,420891124	0,498540876	0,056771	0,132885	0,102513	0,123781	0,036437	0,133384	0,083452	0,093019
Sin Podar										
Error	0,46445		0,133621		0,13038775		0,0864		0,0959	
Desviación	0,019822921		0,022549743		0,0107759		0,027349989		0,009442595	
Intervalos	0,425597076	0,503302924	0,089424	0,177818	0,124048	0,136728	0,032839	0,140051	0,090379	0,10149

Vote										
Algoritmo	50 instancias entrenamiento		Hold out		Hold out repetido (4)		10-XV		4x 10-XV	
J48										
Error	0,41558		0,027027		0,05574325		0,0367		0,0351	
Desviación	0,025116521		0,013466828		0,022324577		0,034235439		0,002207411	
Intervalos	0,366351618	0,464808382	0,000632	0,053422	0,042609	0,068878	-0,03042	0,103782	0,03381	0,036407
Sin Podar										
Error	0,05974		0,027027		0,052365		0,0368		0,0403	
Desviación	0,012078868		0,013466828		0,017770135		0,034424408		0,002976928	
Intervalos	0,036065419	0,083414581	0,000632	0,053422	0,04191	0,06282	-0,03069	0,104258	0,03855	0,042053

11. Discusión resultados

Los algoritmos que nos han dado mayor tasa de error son el 50T instancias, seguido por el de Hold Out y el de validación cruzada que es algo menor.

Esto es algo razonable puesto que, al entrenar con menos instancias, crea un peor modelo y por tanto mayor tasa de error, pero la desviación no es mayor puesto que el conjunto de prueba mantiene todas las instancias.

También cabe observar que, aunque las diferencias de tasa de error entre Hold Out y validación cruzada es pequeña, la variabilidad del Hold Out es mucho más grande, se puede observar muy bien en el Hold Out repetido que presenta mayor desviación y mayores intervalos respecto a la validación cruzada, condicionado a la partición aleatoria.

Y por último validación cruzada son los que presentan las tasas de error más bajas, ya que entrenan con muchos más datos. Ya que son menos variables y los intervalos son más pequeños.

12. Preguntas comparativas sobre validación cruzada repetida

- ¿Qué tasa de error se obtendría con el método 2?

  Se debería obtener la misma tasa de error.

- ¿Cómo espera que varíe la estimación de la varianza con el método 2 frente al método 1?

  En la estimación de la varianza se espera que el valor de esta sea inferior.

- ¿Y los intervalos de confianza?

  Al esperarse que la varianza sea más pequeña, la desviación típica también lo será y como los intervalos depende estas dos, serán a su vez más pequeño.

13. Referencias

- Teodoro Calonge Cano and Carlos Javier Alonso González. Técnicas de Aprendizaje Automático, 2021/22.
- Soybean Data Set. https://archive.ics.uci.edu/ml/datasets/Soybean+(Large)
- Vote Data Set. https://archive.ics.uci.edu/ml/datasets/congressional+voting+records