ENTREGA 4.
PRÁCTICA METODOLOGÍA EXPERIMENTAL: CREACIÓN Y EVALUACIÓN DE HIPÓTESIS Y PODA DE ÁRBOLES.

Patricia Aguado Labrador

Conjuntos de datos:

Soybean: conjunto de datos formado por 35 atributos que contiene información acerca del estado de la soja y las características del entorno de esta, utilizado para diagnosticar enfermedades de esta planta.

Vote: conjunto de datos formado por 17 atributos que contiene información acerca de los votos de cada uno de los congresistas que conforma la Cámara de Representantes de EEUU. La CQA enumera nueve tipos diferentes de votos.

Thoracic_surgery: conjunto de datos formado por 17 atributos que contiene información recopilada de pacientes que se sometieron a extirpaciones pulmonares por cáncer de pulmón primario entre los años 2007 y 2011.

Descripción de los algoritmos:

Para este trabajo utilizamos Weka, en especial los algoritmos J48 con y sin poda y JRIP para la realización de diferentes experimentos.

Ejercicio previo: experimentos con muestras de 50 instancias de los datasets para test

Soybean_50 --> training = 100 - 7.320644217 = 92.68

Vote_50 --> training = 100 - 11.494252874 = 88.51

Thoracic_surgery_50 --> training = 100 - 10.638297872 = 89.36

Datos	**Algoritmo**	**Método: 50 Test, resto**
		Tasa error
Soybean_50	J48	0.04
	Sin podar	0.04
Vote_50	J48	0.04
	Sin podar	0.04
Thoracic_surgery_50	J48	0.22
	Sin podar	0.28

Experimentos de hold out sin repetición

Datos	**Algoritmo**	**Método: Hold out**
		Tasa error
Soybean_50	J48	0.095
	Sin podar	0.13
Vote_50	J48	0.027
	Sin podar	0.027
Thoracic_surgery_50	J48	0.17
	Sin podar	0.25

Experimentos de hold out con repetición

Datos	Algoritmo	Tasa de error		
		2	**3**	**4**
Soybean	J48	0,11	0,11	0,14
	Sin podar	0,11	0,13	0,12
Vote_50	J48	0,08	0,05	0,06
	Sin podar	0,07	0,05	0,06
Thoracic_surgery_50	J48	0,16	0,16	0,15
	Sin podar	0,22	0,22	0,26

Datos	Algoritmo	Método: Hold out repetido			
		Tasa error	**Desviación estándar**	**Intervalos**	
Soybean	J48	0,12	0,0123	0,0994	0,1406
	Sin podar	0,12023	0,0071	0,1083	0,1322
Vote	J48	0,063	0,0108	0,0451	0,0815
	Sin podar	0,06	0,0071	0,0481	0,0719
Thoracic_surgery	J48	0,156	0,0041	0,1498	0,1635
	Sin podar	0,23	0,0163	0,2058	0,2609

Experimentos de validación cruzada sin repetición

Datos	Algoritmo	Método: 10 XV			
		Tasa error	**Desviación estándar**	**Intervalos**	
Soybean	J48	0,08	0,0106	0,0773	0,0827
	Sin podar	0,09	0,0095	0,0876	0,0924
Vote	J48	0,03	0,0057	0,0285	0,0315
	Sin podar	0,04	0,0088	0,0378	0,0423
Thoracic_surgery	J48	0,16	0,0103	0,1574	0,1626
	JRIP	0,23	0,0344	0,2213	0,2388

Experimentos de validación cruzada con repetición

Datos	Algoritmo	Método: 10 XV		
		2	**3**	**4**
Soybean	J48	0,09	0,09	0,08
	Sin podar	0,10	0,10	0,09
Vote	J48	0,03	0,04	0,04
	Sin podar	0,04	0,04	0,04
Thoracic_surgery	J48	0,15	0,15	0,15
	Sin podar	0,22	0,23	0,23

Datos	Algoritmo	Método: 10 XV			
		Tasa error	**Desviación estándar**	**Intervalos**	
Soybean	J48	0,09	0,0068	0,0883	0,0917
	Sin podar	0,1	0,0065	0,0983	0,1017
Vote	J48	0,04	0,0051	0,0387	0,0413
	Sin podar	0,05	0,0076	0,0481	0,0519
Thoracic_surgery	J48	0,16	0,0072	0,1582	0,1618
	JRIP	0,23	0,0036	0,2291	0,2309

Tablas comparativas y discusión de resultados

Soybean

Error

Algoritmo	Hold out	Hold out repetido (4)	10-XV	4 x 10-XV
J48	0,095	0,12	0,082	0,089
Sin podar	0,134	0,12	0,092	0,098

Desviación estándar

Algoritmo	Hold out	Hold out repetido (4)	10-XV	4 x 10-XV
J48	0,019	0,012	0,011	
Sin podar	0,022	0,007	0,009	

Vote

Error

Algoritmo	Hold out	Hold out repetido (4)	10-XV	4 x 10-XV
J48	0,027	0,063	0,034	0,035
Sin podar	0,027	0,06	0,042	0,041

Desviación estándar

Algoritmo	Hold out	Hold out repetido (4)	10-XV	4 x 10-XV
J48	0,013	0,011	0,006	
Sin podar	0,013	0,007	0,009	

Thoracic_surgery

Error

Algoritmo	Hold out	Hold out repetido (4)	10-XV	4 x 10-XV
J48	0,169	0,156	0,155	0,152
Sin podar	0,25	0,233	0,229	0,228

Desviación estándar

Algoritmo	Hold out	Hold out repetido (4)	10-XV	4 x 10-XV
J48	0,029	0,004	0,01	
Sin podar	0,034	0,016	0,034	

1. Soybean:

Con el experimento de validación cruzada de 10 particiones obtenemos el mayor rendimiento tanto con J48 con poda como sin ella. También observamos que la desviación estándar es muy pequeña por lo que podemos asegurar que los datos se encuentran agrupados entorno al valor esperado.

El tamaño del árbol para este conjunto de datos es:

Con poda: 93

Sin poda: 207

2. Votes:
Sin contar el hold-out del ejercicio previo obtenemos el mayor rendimiento en el experimento de validación cruzada de 10 particiones, para el que obtenemos una desviación estándar muy proxima a cero.
El tamaño del árbol para este conjunto de datos es:

Con poda: 11
Sin poda: 71

3. Thoracic_surgery:
En el caso de este dataset obtenemos el mejor rendimiento con un validación cruzada con semilla 2 o 4 (obtenemos la misma tasa de error). También obtenemos un buen resultado con el algoritmo JRIP que realiza una poda incremental para reducir el error. La menor desviación estándar y mayor agrupación de los datos la obtenemos con con el experimento de validación cruzada de 10 particiones con el uso del algoritmo J48 con poda.
El tamaño del árbol para este conjunto de datos es:

Con poda: 1
Sin poda: 157

Referencias
Apuntes vistos en clase de teoría y de laboratorio