PRÁCTICA 1:

MICROARRAY DE EXPRESIÓN GENÉTICA

MINERÍA DE DATOS

MIGUEL CHAVEINTE GARCÍA

DESCRIPCIÓN CONJUNTO DE DATOS

Conjunto de datos sobre la leucemia formado por 72 instancias con 7129 atributos y una clase con dos posibles valores: AML Y ALL.

Realizamos una transformación previa en la que normalizamos los valores de los atributos.

Posteriormente aplicamos los siguientes clasificadores y observamos su tasa de error:

Tasa de error validación cruzada 10 particiones	
J48	0,2083
NB	0,0138
IBK1	0,1527
Regresión Logística	0,125
MLP (H10)	0,0277
SVM (Lineal)	0,0138

J48 es el algoritmo que peor se comporta. Esto puede ser debido al gran número de atributos.

SELECCIÓN DE ATRIBUTOS: FILTRO

Vemos como que los atributos se comparten entre la elección de 4,8,16 y 32 atributos, lo que significa que son relevantes ya que los diferentes métodos de filtro los han seleccionado.
El método CFsubsetEval no termina y esto podría deberse a que hay demasiados atributos en el dataset.

4 atributos:

Incertidumbre Simétrica: 1834, 4847, 1882, 3252, 2288 760, 6041, 6855.

ReliefF: 3252, 4196, 1779, 4847, 2402, 4951, 1834, 1829.
SVM(lineal): 1882, 1834, 1779, 1796, 4196, 5348, 5094, 804.
CFsubsetEval: no terminó.

8 atributos:

Incertidumbre Simétrica: 1834, 4847, 1882, 3252, 2288 760, 6041, 6855.

ReliefF: 3252, 4196, 1779, 4847, 2402, 4951, 1834, 1829.
SVM(lineal): 1882, 1834, 1779, 1796, 4196, 5348, 5094, 804.
CFsubsetEval: no terminó.

16 atributos:

Incertidumbre Simétrica: 1834, 4847, 1882, 3252, 2288, 760, 6041, 6855, 1685, 6376, 2354, 4373, 4377, 4366, 2402, 758.

ReliefF: 3252, 4196, 1779, 4847, 2402, 4951, 1834, 1829, 6041, 2288, 1882, 6201, 1745, 3320, 6919, 2363.

SVM(lineal): 1882, 1834, 1779, 1796, 4196, 5348, 5094, 804, 5107, 3847, 4951, 4847, 2354, 6539, 1933, 2288.

CFsubsetEval: no terminó.

32 atributos:

Incertidumbre Simétrica: 1834, 4847, 1882, 3252, 2288, 760, 6041, 6855, 1685, 6376, 2354, 4373, 4377, 4366, 2402 758, 4328, 1144, 3320, 2642, 2335, 1829, 2128, 6281, 4229, 2020, 1779, 2121, 4196, 1902, 1926, 1400.

ReliefF: 3252, 4196, 1779, 4847, 2402, 4951, 1834, 1829, 6041, 2288, 1882, 6201, 1745, 3320, 6919, 2363, 2111, 4052, 2642, 2121, 1674, 6225, 461, 1249, 4366, 2354, 2020, 6539, 1291, 2546, 1260, 235.

SVM(lineal): 1882, 1834, 1779, 1796, 4196, 5348, 5094, 804, 5107, 3847, 4951, 4847,
2354, 6539, 1933, 2288, 6041, 2300, 3252, 129, 6154, 1941, 2475, 6225, 3320, 3714, 2111, 134, 1975, 6184, 461, 1685.

CFsubsetEval: no terminó.

Tasas de error validación cruzada 10 particiones con selección de atributos				
	Incertidumbre simétrica	**ReliefF**	**Eliminación recursiva con SVM**	**CFsubsetEval**
4 **J48**	0,0972	0,0833	0,0833	x
8 **J48**	0,1527	0,1111	0,0833	x
16 **J48**	0,1527	0,1527	0,125	x
32 **J48**	0,1388	0,1666	0,1527	x

El J48 funciona mejor cuanto menor es el número de atributos significativos seleccionados. De los tres métodos el que mejor funciona con J48 es el de la selección de atributos con SVM.

Tasas de error validación cruzada 10 particiones con selección de atributos				
	Incertidumbre simétrica	**ReliefF**	**Eliminación recursiva con SVM**	**CFsubsetEval**
4 **NB**	0,0555	0,0833	0,0277	x
8 **NB**	0,0555	0,0277	0,0277	x
16 **NB**	0,0416	0,0555	0	x
32 **NB**	0,0416	0,0416	0,0138	x

El algoritmo de Naive Bayes observamos que las tasas de error mejoran al utilizar los métodos de incertidumbre y SVM con mayor número de atributos.

Tasas de error validación cruzada 10 particiones con selección de atributos				
	Incertidumbre simétrica	**ReliefF**	**Eliminación recursiva con SVM**	**CFsubsetEval**
4 **IBK1**	0,0833	0,1111	0	x
8 **IBK1**	0,0694	0,0555	0,0138	x
16 **IBK1**	0,0416	0,0694	0	x
32 **IBK1**	0,0416	0,0694	0	x

El algoritmo de k vecinos funciona muy bien con el algoritmo SVM y con incertidumbre simétrica con bastantes atributos. Con un menor número de atributos funciona bien en el caso de selección de atributos.

Tasas de error validación cruzada 10 particiones con selección de atributos				
	Incertidumbre simétrica	**ReliefF**	**Eliminación recursiva con SVM**	**CFsubsetEval**
4 **R.LOGIST**	0,0694	0,0555	0	x
8 **R.LOGIST**	0,0555	0,0972	0,0138	x
16 **R.LOGIST**	0,0417	0,0694	0	x
32 **R.LOGIST**	0,0416	0,0416	0	x

En el algoritmo de regresión la selección de atributos funciona muy bien a mayor es el número de atributos seleccionados.

Tasas de error validación cruzada 10 particiones con selección de atributos				
	Incertidumbre simétrica	**ReliefF**	**Eliminación recursiva con SVM**	**CFsubsetEval**
4 **MLP (H10)**	0,0694	0,0555	0	x
8 **MLP (H10)**	0,0416	0,0694	0	x
16 **MLP (H10)**	0,0138	0,0694	0	x
32 **MLP (H10)**	0,0277	0,0277	0	x

El algoritmo de MLP con 10 capas ocultas funciona muy bien con el algoritmo SVM y con el resto a mayor número de atributos funcionan mejor.

Tasas de error validación cruzada 10 particiones con selección de atributos				
	Incertidumbre simétrica	**ReliefF**	**Eliminación recursiva con SVM**	**CFsubsetEval**
4 **SVM (lineal)**	0,0694	0,0555	0	x
8 **SVM (lineal)**	0,0694	0,0555	0	x
16 **SVM (lineal)**	0,0555	0,0277	0	x
32 **SVM (lineal)**	0,0277	0,0277	0	x

Al utilizar el algoritmo SVM para clasificar vemos que las tasas de error son menores al utilizar el método de incertidumbre simétrica y ReliefF con un número de atributos alto, y que con eliminación recursiva obtiene las mejores tasas de error.

DISCUSIÓN DE RESULTADOS:

En todos los métodos de filtros utilizados notamos una mejoría en las tasas de error, pero podemos observar que mediante el filtro de SVM mejora aún más los resultados.

SELECCIÓN DE ATRIBUTOS: ENVOLTORIO

Atributos seleccionados:

J48: 4847.

NB: 6, 461, 760, 6615.

IBK1: 28, 1834, 3258, 3549.

Reg. Log: 43, 1882, 6049.

MLP(10): 1795, 1834, 2288.

SVM(lineal): 162, 1796, 2111, 3252.

Tasas de error validación cruzada 10 particiones						
	J48	**NB**	**IBK1**	**Reg. Log**	**MLP(10)**	**SVM(lineal)**
J48	0,0555					
NB		0,0138				
IBK1			0			
Reg. Log				0,0138		
MLP(10)					0,0555	
SVM(lineal)						0,0277

Los peores algoritmos son el J48 y MLP, aunque las tasas de error son muy pequeñas en ambos casos (5.5%).

El que mejor se comporta es k vecinos.

COMPARACIÓN CON MÉTODOS DE FILTRO:

Los atributos seleccionados son en mayoría los que habían sido seleccionados en los métodos de filtro, como el 1834,4847,760, …

Tasas de error:
Con el método de J48 obtenemos una tasa de error bastante menor que cuando empleábamos este algoritmo junto los métodos de filtro para la clasificación del conjunto de datos.

En el resto de los algoritmos obtenemos tasas de error pequeña en los métodos de envoltorio, al igual que ocurría con el SVM en la selección de atributos (con tasas de error nulas en algunos casos)

PCA

El algoritmo con menor tasa de error era el IBK1. Con el número de atributos significativos n=4.

No he podido realizar las ejecuciones de los ejercicios planteados ya que el análisis de componentes principales no me llegó a ejecutar, incluso dejándolo, ejecutándolo varias horas.

CONCLUSIONES

Los métodos envoltorio nos dan mejores tasas de error, incluso muy buenas, aunque esto nos lleva a algoritmos con sobreajuste si utilizan el método de aprendizaje como evaluador.

La eliminación recursiva con SVM junto a los métodos de filtro nos ha dado las mejores tasas de error con cualquier tipo de algoritmos junto a la validación cruzada.