PRÁCTICA SOBRE LA COMPARACIÓN DE ALGORITMOS DE APRENDIZAJE

Práctica 2

Resumen ejecutivo de los conjunto de datos

Descripción del conjunto de datos

Los 9 primeros conjuntos de datos han sido obtenidos de la carpeta de datasets de weka.

Los 3 conjuntos de datos restantes se han obtenido del repositorio de la UCI: http://archive.ics.uci.edu/ml/in dex.html

1) Conjunto de datos: Soybean.arff

El conjunto de datos tiene 683 instancias y 36 atributos (35 atributos + 1 clase), todos son de tipo Nominal y una clase de tipo Nominal (class), donde la clase puede tomar 19 posibles valores: {diaporthe-stem-canker, charcoal-rot, rhizoctonia-root-rot, phytophthora-rot, brown-stem-rot, powdery-mildew, downy-mildew, brown-spot, bacterial-blight, bacterial-pustule, purple-seed-stain, anthracnose, phyllosticta-leaf-spot, alternarialeaf-spot, frog-eye-leaf-spot, diaporthe-pod-&-stem-blight, cyst-nematode, 2-4-d-injury, herbicide-injury}.

La distribución de la clase en porcentajes es de un 2,92% para diaporthe-stem-canker, charcoal-rot, rhizoctonia-root-rot, powdery-mildew, downy-mildew, bacterial-blight, bacterial-pustule, purple-seed-stain y phyllosticta-leaf-spot. Un 6,44% para brown-stem-rot, anthracnose. Un 12,88% para phytophthora-rot. Un 13,46% para brown-spot, y para alternarialeaf-spot, frog-eye-leaf-spot un 13,32%. Y por último las que tienen una menor distribución 1,17% para herbicide-injury, 2,04% para cyst-nematode, 2,19% y 2,34% para diaporthe-pod-&-stem-blight y 2-4-d-injury respectivamente.

Arff transformado → Scaled : soybean-scaled.arff (No sería necesario puesto que todos los atributos son nominales, y realizar el escalado no modifica el dataset)

2) Conjunto de datos: vote.arff

El conjunto de datos tiene 435 instancias y 17 atributos (17 atributos + 1 clase), todos son de tipo Nominal, y los 17 pueden tomar el valor 'y' (yes) o 'n' (no), y una clase de tipo Nominal (Class), donde la clase puede tomar **2 posibles valores**: democrat y republican.

La distribución de la clase en porcentajes es de un 54,25% para democrat, y 38,62% para republican.

Arff transformado → Scaled y eliminación las instancias Missings: vote-scaled.arff
Primero eliminamos los Missings ya que los contiene el dataset, para ello aplicamos el filtro → unsupervised → attributes → ReplaceMissingValue, y como vemos todos los atributos Missings que había han desaparecido, y despues aplicamos el escalado → Normalize.

3) Conjunto de datos: labor.arff

El conjunto de datos tiene 57 instancias y 17 atributos (17 atributos + 1 clase), todos son de tipo Nominal y Numeric (duration,wage-increase-first-year,wage-increase-third-year,working-hours,standby-pay,shift-differential,statutory-holidays → TIPO NUMERICO, y cost-of-living-adjustment, pension,education-allowance, vacation, longterm-disability-assistance,contribution-to-dental-plan,bereavement-assistance,contribution-to-health-plan → TIPO NOMINAL) y una clase de tipo Nominal (class), donde la clase puede tomar 2 posibles valores: bad y good.
La distribución de la clase en porcentajes es de un 35,09% para bad, y 64,91% para good.

Arff transformado → Scaled y eliminación las instancias Missings: labor-scaled.arff
Primero eliminamos los Missings ya que los contiene el dataset, para ello aplicamos el filtro → unsupervised → attributes → ReplaceMissingValue, y como vemos todos los atributos Missings que había han desaparecido, y después aplicamos el escalado → Normalize.

4) Conjunto de datos: ionosphere.arff

El conjunto de datos está formado por **351 instancias** y **35 atributos,** todos ellos de tipo numérico. La clase es de tipo nominal y puede tomar **2 posibles valores**: b y g. En cuanto a la distribución de cada valor: b (35,9 %) y g (64,1 %).

El conjunto de datos no presenta valores ausentes y tampoco outliers.

Pasamos a normalizar el conjunto de datos escalando en el intervalo [0,1] para darles a todos los atributos el mismo peso. Llamamos al conjunto resultante **ionosphere-scaled.arff,** que será con el que trabajaremos a partir de ahora.

5) Conjunto de datos: diabetes.arff

El conjunto de datos está formado por **768 instancias** y **9 atributos,** todos ellos de tipo numérico. La clase es de tipo nominal y puede tomar **2 posibles valores**: tested_negative y tested_positive. En cuanto a la distribución de cada valor: tested_negative (65,1 %) y tested_positive (34,9 %).

El conjunto de datos no presenta valores ausentes y tampoco outliers.

Pasamos a normalizar el conjunto de datos escalando en el intervalo [0,1] para darles a todos los atributos el mismo peso. Llamamos al conjunto resultante **diabetes-scaled.arff,** que será con el que trabajaremos a partir de ahora.

6) Conjunto de datos: glass.arff

El conjunto de datos está formado por **214 instancias** y **10 atributos,** todos ellos de tipo numérico. La clase es de tipo nominal y puede tomar **7 posibles valores**: {build wind float, build wind non-float, vehic wind float, vehic wind non-float, containers, tableware, headlamps}. En cuanto a la distribución de cada valor:

- build wind float (32,7 %)
- build wind non-float (35,5 %)
- vehic wind float (7,9 %)
- vehic wind non-float (0 %)
- containers (6,1 %)
- tableware (4,2 %)
- headlamps (13,6 %)

El conjunto de datos no presenta valores ausentes y tampoco outliers.

Pasamos a normalizar el conjunto de datos escalando en el intervalo [0,1] para darles a todos los atributos el mismo peso. Llamamos al conjunto resultante **glass-scaled.arff,** que será con el que trabajaremos a partir de ahora.

7) Conjunto de datos: segment-test.arff

El conjunto de datos está formado por **810 instancias** y **20 atributos,** todos ellos de tipo numérico. La clase es de tipo nominal y puede tomar **7 posibles valores**: {brickface, sky, foliage, cement, window, path, grass}. En cuanto a la distribución de cada valor:

- brickface (15,4 %)
- sky (13,6 %)
- foliage (15,1 %)
- cement t (13,6 %)
- window (15,5 %)
- path (11,6 %)
- grass (15,2 %)

El conjunto de datos no presenta valores ausentes y tampoco outliers.

Pasamos a normalizar el conjunto de datos escalando en el intervalo [0,1] para darles a todos los atributos el mismo peso. Llamamos al conjunto resultante **segment-test-scaled.arff,** que será con el que trabajaremos a partir de ahora.

8) Conjunto de datos: breast-cancer.arff

El conjunto de datos está formado por **286 instancias** y **10 atributos,** todos ellos de tipo nominal. La clase también es de tipo nominal y puede tomar **2 posibles valores**: no-recurrence-events y recurrence-events. En cuanto a la distribución de cada valor: no-recurrence-events (70,3 %) y recurrence-events (29,7 %).

El conjunto de datos sí que presenta algunos valores ausentes. Para eliminarlos aplicamos el filtro correspondiente: **ReplaceMissingValues**. Una vez eliminados, como los atributos son todos nominales, guardamos el conjunto de datos como **breast-cancer-scaled.arff,** que será con el que trabajaremos a partir de ahora.

9) Conjunto de datos: credit-g.arff

El conjunto de datos está formado por **1000 instancias** y **21 atributos,** 13 de ellos de tipo nominal y los otros 7 de tipo numérico. La clase es de tipo nominal y puede tomar **2 posibles valores**: good y bad. En cuanto a la distribución de cada valor: good (70 %) y bad (30 %).

El conjunto de datos no presenta valores ausentes ni outliers.

Pasamos a normalizar el conjunto de datos escalando en el intervalo [0,1] para darles a todos los atributos numéricos el mismo peso. Llamamos al conjunto resultante **credit-g-scaled.arff,** que será con el que trabajaremos a partir de ahora.

10) Conjunto de datos: Iris.arff

El conjunto de datos está formado por **150 instancias** y **5 atributos,** todos ellos de tipo numérico. La clase es de tipo nominal y puede tomar **3 posibles valores**: Iris-setosa, Iris-versicolor e Iris-virginica. En cuanto a la distribución de cada valor: Iris-setosa (33,33 %), Iris-versicolor (33,33 %) e Iris-virginica (33,33 %).

El conjunto de datos no presenta valores ausentes ni outliers.

Pasamos a normalizar el conjunto de datos escalando en el intervalo [0,1] para darles a todos los atributos el mismo peso. Llamamos al conjunto resultante **credit-g-scaled.arff,** que será con el que trabajaremos a partir de ahora.

11) Conjunto de datos: Thoracic-Surgery.arff

El conjunto de datos está formado por **470 instancias** y **17 atributos**, 3 de ellos de tipo numérico y los otros 13 de tipo nominal. La clase es de tipo nominal y puede tomar 2 **posibles valores**: T y F. En cuanto a la distribución de cada valor: T (14,9 %) y F (85,1 %).

El conjunto de datos no presenta valores ausentes ni outliers.

Pasamos a normalizar el conjunto de datos escalando en el intervalo [0,1] para darles a todos los atributos numéricos el mismo peso. Llamamos al conjunto resultante
Thoracic-Surgery-scaled.arff, que será con el que trabajaremos a partir de ahora.

12) Conjunto de datos: Tmusic.arff

El conjunto de datos está formado por **211 instancias** y **19 atributos**, todos ellos de tipo numérico. La clase es de tipo nominal y puede tomar **4 posibles valores:** {R1, R2, R3, R4}. En cuanto a la distribución de cada valor:

- R1 (42,6 %)
- R2 (18,5 %)
- R3 (18,0 %)
- R4 (20,9 %)
-

El conjunto de datos no presenta valores ausentes y tampoco outliers.

Pasamos a normalizar el conjunto de datos escalando en el intervalo [0,1] para darles a todos los atributos el mismo peso. Llamamos al conjunto resultante **Tmusic-scaled.arff,** que será con el que trabajaremos a partir de ahora.

Comparación 2 métodos mismo conjunto de datos

Test de McNemar

Para generar los conjuntos de entrenamiento y prueba utilizamos el filtro no supervisado de instancias: *resample*. Para obtener el **conjunto de prueba** tenemos que marcar invertSelection y noReplacement a **True** y un porcentaje del **66.67 %.** Obtenemos un conjunto de **228 instancias**.

Para el **conjunto de entrenamiento**, marcamos invertSelection como **False** y noReplacement a **True** y un porcentaje del **66.67 %.** Obtenemos un conjunto de **455 instancias.**

Entrenamos los algoritmos A = OneR y B = J48 sobre el conjunto de entrenamiento generando las hipótesis h_A y h_B.

Instancias bien clasificadas por h_A : 69 *Instancias mal clasificadas por h_A : 159*

Instancias bien clasificadas por h_B : 214 *Instancias mal clasificadas por h_B : 14*

[Valores obtenidos mediante un Script realizado en Python]

Tabla de contingencia resultante:

McNemar	Mal clasificados por h_B	Bien clasificados por h_B
Mal clasificados por h_A	12 **n_{00}**	147 **n_{01}**
Bien clasificados por h_A	2 **n_{10}**	67 **n_{11}**

Bajo la hipótesis nula, los 2 algoritmos deben tener la misma tasa de error: $n_{01} = n_{10}$.

Una vez dividido el conjunto de datos inicial y entrenados los algoritmos sobre T, procedemos a aplicar el **Test de McNemar**. Este se basa en un test X^2 para la bondad del ajuste que compara la distribución esperada de la tabla de contingencia con la hipótesis nula con la distribución observada.

Este **test es aplicable** si $n_{01} + n_{10} > 25$.

Como $n_{01} + n_{10} = 2 + 147 = 149 > 25$, el test de McNemar es aplicable.

El estadístico $\frac{(|n01-n10|-1)\text{^}2}{n01+n10}$ = 139.17 se distribuye aproximadamente como una chi-cuadrada con un grado de libertad con una confianza del 95%, que es igual a 3.841459.

Como 139.17 (estadístico) **es mayor que** 3.841459 (X^2) **se rechaza la hipótesis nula con una confianza del 95%, es decir, los 2 clasificadores son significativamente distintos.**

Sin embargo, este test es muy exigente.

- Bajo error Tipo I
- **Alto error Tipo II**

Debido a esto, se suele recurrir a otros mecanismos como los siguientes:

Validación cruzada con test de Student remuestreado pareado (corregido)

Test de Student pareado

Disponemos de un único conjunto de datos D = soybean-scaled.arff. Para realizar el test de Student no dirigimos a Weka-Experimenter y seleccionamos para este caso validación cruzada de 10 particiones con 1 sola repetición. Procedemos a seleccionar los clasificadores correspondientes y ejecutamos. En la pestaña de Analyse comprobamos que el nivel de confianza sea del 5%, que el campo a comparar sea la tasa de acierto y como se trata del test NO corregido, en *Testing with* debemos seleccionar la opción Paired T-Tester. Finalmente pulsamos en Perform test.

Utilizamos como algoritmo base primero NB

Tasa de acierto %			
NB	J48	IB1	SVM
92.08	92.39	91.64	93.85

Ahora, como algoritmo base SVM

Tasa de acierto %			
SVM	J48	IB1	NB
93.85	92.39	91.64	92.08

Test de Student pareado corregido

Mismos pasos que en caso anterior, la única diferencia es que hay que marcar en *Testing with* la opción Paired T-Tester (corrected), que en Weka es la opción que viene marcada por defecto.

Utilizamos como algoritmo base primero NB

Tasa de acierto %			
NB	J48	IB1	SVM
92.08	92.39	91.64	93.85

Ahora, como algoritmo base SVM

Tasa de acierto %			
SVM	J48	IB1	NB
93.85	92.39	91.64	92.08

Breve discusión de los resultados

Para **el test de Student pareado** con NB como clasificador base se observa que obtienen mayores tasas de acierto tanto J48 como SVM. El clasificador IB1 es el que menor tasa de acierto presenta.

Con SVM como algoritmo base tenemos que todas las tasas de error del resto de clasificadores son peores.

Para **el test de Student pareado corregido** obtenemos los mismos resultados.

Validación cruzada con repetición con test de Student remuestreado pareado (corregido)

Test de Student pareado

Disponemos de un único conjunto de datos D = soybean-scaled.arff. Para realizar el test de Student no dirigimos a Weka-Experimenter y seleccionamos para este caso validación cruzada de 10 particiones con 10 repeticiones. Procedemos a seleccionar los clasificadores correspondientes y ejecutamos. En la pestaña de Analyse comprobamos que el nivel de confianza sea del 5%, que el campo a comparar sea la tasa de acierto y como se trata del test NO corregido, en *Testing with* debemos seleccionar la opción Paired T-Tester. Finalmente pulsamos en Perform test.

Utilizamos como algoritmo base primero NB

Tasa de acierto %			
NB	**J48**	**IB1**	**SVM**
92.20	92.63	91.35	93.10

Ahora, como algoritmo base SVM

Tasa de acierto %			
SVM	**J48**	**IB1**	**NB**
93.10	92.63	91.35	92.20

Test de Student pareado corregido

Mismos pasos que en caso anterior, la única diferencia es que hay que marcar en *Testing with* la opción Paired T-Tester (corrected), que en Weka es la opción que viene marcada por defecto.

Utilizamos como algoritmo base primero NB

Tasa de acierto %			
NB	J48	IB1	SVM
92.20	92.63	91.35	93.10

Ahora, como algoritmo base SVM

Tasa de acierto %			
SVM	J48	IB1	NB
93.10	92.63	91.35	92.20

Breve discusión de los resultados

Observamos unos resultados similares a los obtenidos en la validación cruzada sin repetición. Hay que destacar que el clasificador con el que se obtienen mayores tasas de acierto es **SVM**.

Dos métodos, varios conjuntos de datos

Test de signos

Comparar OneR y J48 sobre los 12 conjuntos de datos, estimando su tasa de error mediante validación cruzada con 10 particiones

Dos métodos y varios conjuntos de datos.

Abrimos Weka-Experimenter, empleamos validación cruzada con 10 particiones. Seleccionamos los 12 conjuntos de datos y los 2 métodos: OneR y J48.

Tasa de error % (α = 0.05)		
	OneR	**J48**
Soybean	66.47	7.61
Vote	4.36	3.67
Labor	28.00	18.33
Ionosphere	19.08	8.54
Diabetes	28.52	26.03
Glass	41.99	33.25
Segment-test	36.17	6.54
Breast-cancer	34.26	24.46
Credit-g	33.90	29.20
Iris	8.00	4.00
Thoracic-Surgery	16.60	15.53
Tmusic	26.13	18.51

Victorias:	0	12

Comparar J48 y NB sobre los 12 conjuntos de datos, estimando su tasa de error mediante validación cruzada con 10 particiones

Tasa de error % (α = 0.05)		
	J48	**NB**
Soybean	7.61	7.92
Ionosphere	3.67	9.86 *
Vote	18.33	12.00
Diabetes	8.54	17.38 *
Labor	26.03	23.69
Glass	33.25	50.48 *
Segment-test	6.54	13.21 *
Breast-cancer	24.46	27.94
Credit-g	29.20	24.40
Iris	4.00	5.33
Thoracic-Surgery	15.53	22.13
Segment-challenge	18.51	42.23 *

Victorias:	9	3

Como la significancia es del 5%, rechazamos la hipótesis nula si el número de victorias es superior a:

$$N/2 + 1.96 * \sqrt{\frac{N}{2}}$$

Como tenemos 12 conjuntos de datos N=12 y el resultado que se obtiene es = 10.80.

- En la primera comparación entre OneR y J48, este último obtiene 12 victorias lo cual es superior al valor calculado. Por esto motivo, se rechaza la hipótesis nula y podemos asegurar que el algoritmo J48 es mejor.

- Para la segunda comparación entre J48 y NB, el primero obtiene 9 victorias por 3 del segundo. Pero podemos ver que las 9 victorias no son superiores al valor calculado por lo que acepta la hipótesis nula. No podemos asegurar que ninguno de los algoritmos de clasificación sea mejor que el otro.

Rankings

Varios métodos, varios conjuntos de datos.

Abrimos Weka-Experimenter, empleamos validación cruzada repetida de 5 particiones con 5 repeticiones. Seleccionamos los 12 conjuntos de datos y los 5 algoritmos de clasificación: J48, OneR, IB3, NB, SVM kernel lineal (SMO).

	Tasa de error %					Rankings				
	J48	**OneR**	**3NN**	**NB**	**SVM**	**J48**	**OneR**	**3NN**	**NB**	**SVM**
Soybean	7,64	66,35	8,46	8,02	6,94	2	5	4	3	1
Vote	3,54	4,37	6,07	9,79	4,28	1	3	4	5	2
Labor	17,85	23,79	11,55	8,73	9,12	4	5	3	1	2
Ionosphere	10,94	18,80	13,67	17,67	11,91	1	5	3	4	2
Diabetes	27,39	27,81	26,28	24,58	22,99	4	5	3	2	1
Glass	33,07	44,95	32,05	50,58	42,89	2	4	1	5	3
Segment-test	6,12	39,06	7,85	13,56	7,78	1	5	3	4	2
Breast-cancer	25,39	32,59	27,00	26,71	30,70	1	5	3	2	4
Credit-g	27,98	33,50	28,04	24,90	24,50	3	5	4	2	1
Iris	5,33	6,93	5,20	4,93	3,60	4	5	3	2	1
Thoracic-Surgery	15,32	16,30	17,32	25,28	15,49	1	3	4	5	2
Tmusic	20,00	30,34	41,87	41,98	36,87	1	2	4	5	3
Ranking promedio:						2,08	4,33	3,25	3,22	2

Test de Friedman sobre el conjunto de datos

Hipótesis nula: los 5 métodos son equivalentes.

El estadístico de Friedman se comporta como una **distribución X^2 con k-2 grados de libertad**.

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]$$

N = 12 conjuntos de datos

k = 5 algoritmos

$\mathbf{X^2_F} = \frac{12*12}{5*6}\left[(2{,}08^2 + 4{,}33^2 + 3{,}25^2 + 3{,}22^2 + 2^2) - \frac{5*6^2}{4}\right] = 14{,}43$

Valor crítico X^2, con 3 grados de libertad, α = 0,05 ➔ 7,815

Como $X^2_F > X^2$, se rechaza la hipótesis nula: **los rankings son significativamente distintos**.

Test de Iman y Davenport sobre el conjunto de datos

El estadístico de Imán y Davenport se comporta según la distribución F con k-1 y (k-1)(N-1) grados de libertad.

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1) - \chi_F^2}$$

N = 12

k = 5

$X^2_F = 14{,}43$

$\mathbf{F_F} = \frac{11*14{,}43}{12*4-14{,}43} = 4{,}73$

Valor crítico F, con 4 y 44 grados de libertad, α = 0,05 ➔ 2,584

En este caso, como $F_F > F$ también rechazamos la hipótesis nula: **los rankings son significativamente distintos**.

Test post-hoc

Como hemos comprobado, los rankings son significativamente diferentes, es por ello que realizamos el test Post-hoc.

Test de Nemenyi sobre el conjunto de datos

Dos métodos son significativamente diferentes si sus rankings promedios difieren al menos en la distancia crítica de Nemenyi, $CD = q_\alpha\sqrt{\frac{k(k-1)}{6N}}$

N = 12

k = 5

$q_{0,05}$ = 2,728 para 5 clasificadores

CD = 2,728 * $\sqrt{\frac{5*4}{6*12}}$ = **1,438**

Recordamos los **rankings promedios** de cada clasificador:

J48	OneR	3NN	NB	SVM
2,08	4,33	3,25	3,22	2

Realizando las restas podemos comprobar que los métodos significativamente diferentes son:

- OneR y SVM ➔ 4,33 – 2 = 2,33 > 1,438
- OneR y J48 ➔ 4,33 – 2,08 = 2,25 > 1,438

Test de Bonferroni-Dunn sobre el conjunto de datos

Test con menos comparaciones que el anterior ya que sólo se compara un método frente a los demás.

La distancia crítica tiene la misma expresión que la de Nemenyi, únicamente cambia el valor de q_α.

Realizamos 2 test de Bonferroni-Dunn:

1. OneR frente al resto (peor frente al resto)

   CD = 2,498 * $\sqrt{\frac{5*4}{6*12}}$ = 1.22

   4.33 – 1.22 = **3.11**

2. SVM frente al resto (mejor frente al resto)

   CD = 2,498 * $\sqrt{\frac{5*4}{6*12}}$ = 1.22

   2 – 1.22 = **0.78**