MA '20
EB 12

Algoritmos de Estimación de Distribuciones para la Selección Simultánea de Instancias y Atributos

MAEB 2012 – Albacete

8-10 Febrero

Pablo Bermejo, José A. Gámez, **Ana M. Martínez** y José M. Puerta

Universidad de Castilla-La Mancha

Resumen

1. Pre-procesamiento de bases de datos
2. IFS - Propuestas
3. Experimentos
4. Conclusiones y Trabajo Futuro

1. Pre-procesamiento de bases de datos

- Las técnicas de pre-procesamiento mejoran la calidad de los datos con los que se pretende construir modelos predictivos:

Instance ID	X_1	X_2	$\dots$	X_n	C
1	x_{11}	x_{12}	$\dots$	x_{1n}	c_1
2	x_{21}	x_{22}	$\dots$	x_{2n}	c_2
3	x_{31}	x_{32}	$\dots$	x_{3n}	c_3
4	x_{41}	x_{42}	$\dots$	x_{4n}	c_4
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
N	x_{N1}	x_{N2}	$\dots$	x_{Nn}	c_N

- Cuando ***n*** es del orden de miles, hablamos de bases de datos de **alta dimensionalidad**. Si ***N*** es muy grande, bases de datos masivas.
- Otro problema común: *imbalanceado*.

1. Pre-procesamiento de bases de datos

- Las técnicas de pre-procesamiento mejoran la calidad de los datos con los que se pretende construir modelos predictivos:

Selección de atributos(FS)

Instance ID	X_1	X_2	$\ldots$	X_n	C
1	x_{11}	x_{12}	$\ldots$	x_{1n}	c_1
2	x_{21}	x_{22}	$\ldots$	x_{2n}	c_2
3	x_{31}	x_{32}	$\ldots$	x_{3n}	c_3
4	x_{41}	x_{42}	$\ldots$	x_{4n}	c_4
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
N	x_{N1}	x_{N2}	$\ldots$	x_{Nn}	c_N

- Cuando ***n*** es del orden de miles, hablamos de bases de datos de **alta dimensionalidad**. Si ***N*** es muy grande, bases de datos masivas.
- Otro problema común: *imbalanceado*.

1. Pre-procesamiento de bases de datos

- Las técnicas de pre-procesamiento mejoran la calidad de los datos con los que se pretende construir modelos predictivos:

Selección de atributos(FS) Construcción de Atributos

Instance ID	X_1	X_2	$\ldots$	X_n	C	X_{n+1}
1	x_{11}	x_{12}	$\ldots$	x_{1n}	c_1	…
2	x_{21}	x_{22}	$\ldots$	x_{2n}	c_2	…
3	x_{31}	x_{32}	$\ldots$	x_{3n}	c_3	…
4	x_{41}	x_{42}	$\ldots$	x_{4n}	c_4	…
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	…
N	x_{N1}	x_{N2}	$\ldots$	x_{Nn}	c_N	…

- Cuando ***n*** es del orden de miles, hablamos de bases de datos de **alta dimensionalidad**. Si ***N*** es muy grande, bases de datos masivas.
- Otro problema común: *imbalanceado*.

1. Pre-procesamiento de bases de datos

- Las técnicas de pre-procesamiento mejoran la calidad de los datos con los que se pretende construir modelos predictivos:

Selección de atributos(FS) Construcción de Atributos

Muestrear instancias

Instance ID	X_1	X_2	$\ldots$	X_n	C	X_{n+1}
1	x_{11}	x_{12}	$\ldots$	x_{1n}	c_1	…
2	x_{21}	x_{22}	$\ldots$	x_{2n}	c_2	…
3	x_{31}	x_{32}	$\ldots$	x_{3n}	c_3	…
4	x_{41}	x_{42}	$\ldots$	x_{4n}	c_4	…
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	…
N	x_{N1}	x_{N2}	$\ldots$	x_{Nn}	c_N	…
N+1	…	…	…	…	…	

- Cuando ***n*** es del orden de miles, hablamos de bases de datos de **alta dimensionalidad**. Si ***N*** es muy grande, bases de datos masivas.
- Otro problema común: *imbalanceado*.

1. Pre-procesamiento de bases de datos

- Las técnicas de pre-procesamiento mejoran la calidad de los datos con los que se pretende construir modelos predictivos:

Selección de atributos(FS) Construcción de Atributos
Muestrear instancias **Seleccionar Instancias (IS)**

Instance ID	X_1	X_2	$\dots$	X_n	C	X_{n+1}
1	x_{11}	x_{12}	$\dots$	x_{1n}	c_1	...
2	x_{21}	x_{22}	$\dots$	x_{2n}	c_2	...
3	x_{31}	x_{32}	$\dots$	x_{3n}	c_3	...
4	x_{41}	x_{42}	$\dots$	x_{4n}	c_4	...
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	...
N	x_{N1}	x_{N2}	$\dots$	x_{Nn}	c_N	...
N+1	...	...	...	...	...	

- Cuando ***n*** es del orden de miles, hablamos de bases de datos de **alta dimensionalidad**. Si ***N*** es muy grande, bases de datos masivas.
- Otro problema común: *imbalanceado*.

1. Pre-procesamiento de bases de datos

- **FS (Selección de atributos):**
  - Reduce la anchura de la base de datos: ligereza, mejores modelos, más entendibles.
  - Las métricas utilizadas para seleccionar un atributo suelen calcularse **a partir de las instancias disponibles.**

- **IS (Selección de instancias):**
  - Reduce la longitud de la base de datos: ligereza, clusters mejor definidos, gran mejora para clasificadores perezosos.
  - Las métricas utilizadas para seleccionar una instancia suelen calcularse **a partir de los atributos disponibles.**

FS influye en IS

IS influye en FS

¿Cuál realizar antes? → Selección Simultánea (IFS)

2. IFS

- **IFS-CHC:** algoritmo evolutivo adaptativo CHC**.**
- **IGA**: algoritmo genético inteligente con operador de cruce ortogonal.
- **HGA:** algoritmo genético híbrido: técnicas de búsqueda local+AG.
- Una de las propuestas más recientes y exitosas es **IFS-CoCo**:
  - Co-evolución genética de 3 poblaciones
    - 1 población para FS
    - 1 población para IS
    - 1 población para IFS
  - Resultados muy buenos
  - Wrapper: caro computacionalmente

EDAs

- Algoritmos evolutivos que trabajan sobre conjuntos de poblaciones de soluciones candidatas.

Algoritmo 1 Algoritmo de estimación de distribuciones (EDA).

$D_0 \leftarrow$ Generar la población inicial (m individuos)
Evaluar la población D_0
$l = 1$
Repeat until condición de parada
  $D_{l-1}^{S_e} \leftarrow$ Seleccionar $s \leq m$ individuos de D_{l-1}
  Estimar un modelo probabilístico $\mathcal{M}$ a partir de $D_{l-1}^{S_e}$
  $D_{l-1}^{m} \leftarrow$ Muestrear m individuos a partir de $\mathcal{M}$
  Evaluar D_{l-1}^{m}
  $D_l \leftarrow$ Seleccionar m individuos de $D_{l-1} \cup D_{l-1}^{m}$
  $l = l + 1$
end

EDAs

- **Ventajas** (frente a los AGs):
  - Menor número de parámetros a ajustar.
  - Mayor expresividad y transparencia del modelo probabilístico que guía el proceso de búsqueda.

- Existen multitud de EDAs:
  - Sin dependencias: **UMDA**.
    - Distribución de probabilidad conjunta a partir de las distribuciones univariadas independientes.
  - Dependencias bivariadas.
  - Dependencias múltiples.

2. IFS

- Proponemos 2 nuevos métodos para realizar IFS, ambos basados en evolución de EDAs (UMDA), y con evaluación principalmente filter:
  - **IFS-EDAig**
  - **IFS-EDAcfs**
  - En ambas búsquedas, las poblaciones se inicializan dando la probabilidad a cada instancia inversamente proporcional a la cardinalidad de su clase, para intentar sesgar el resultado a una base de datos balanceada → IFS + balanceado.

2. IFS

- **IFS-EDAig:**
  - Fase filter (UMDA): repetida *n* veces
    - Un individuo representa explícitamente las instancias seleccionadas
    - Los atributos seleccionadas se representan implícitamente con un ranking de los *k* mejores atributos por IG
    - La suma de estos *k* atributos es la bondad de un individuo
  - Fase wrapper: comparación del mejor individuo de cada UMDA$_k$

Algoritmo IFS-EDAig.

In Base de datos $\mathcal{T}$, nº atributos n, nº instancias l, tamaño población m, nº generaciones x, tamaño población aprendizaje s, clasificador $\mathcal{C}$

Out Base de datos reducida $\mathcal{R}_{\mathcal{T}}$

for $k \leftarrow 1$ **to** n
//fase filter
  mejorInd $\leftarrow$ UMDA$_{ig}(k,l,m,s,x,\mathcal{T})$ $\longrightarrow$ $\boxed{f(c[]) = \sum_{i=1}^{k} IG_{T_{c[]}}(x_i)}$
// fase wrapper
  $\mathbf{T}'^{\downarrow\mathbf{k}} \leftarrow$ proyectar$(\mathcal{T}'$,mejorInd$)$ $\longrightarrow$ K mejores atributos: ranking por IG
  tasaAciertos $\leftarrow$ crossValidate$(\mathbf{T}'^{\downarrow\mathbf{k}}, \mathcal{C})$
  if(tasaAciertos > mejorTasaAc)
    mejorTasaAc$\leftarrow$ tasaAciertos
    mejor$\mathcal{T} \leftarrow \mathbf{T}'^{\downarrow\mathbf{k}}$
end
return mejor$\mathcal{T}$

2. IFS

- **IFS-EDAcfs:** filter puro: 1 sola búsqueda UMDA
  - Un individuo representa explícitamente las instancias seleccionadas.
  - Los atributos seleccionadas se representan implícitamente con una búsqueda voraz con métrica CFS (correlation-based FS).
  - La bondad de un individuo es la devuelta por dicha búsqueda.

Algoritmo IFS-EDAcfs.

In Base de datos $\mathcal{T}$, nº instancias l, tamaño población m, nº generaciones x, tamaño población aprendizaje s

Out Base de datos reducida $\mathcal{R}_{\mathcal{T}}$

//fase filter

mejorInd $\leftarrow$ $\text{UMDA}_{cfs}(l, m, s, x, \mathcal{T})$

return $\mathbf{T}'^{\downarrow \mathbf{k}} \leftarrow \text{proyectar}(\mathcal{T}', \text{mejorInd})$

3. Experimentos

- Utilizamos el siguiente corpus para la evaluación de nuestras propuestas y comparación con otros métodos:

Nombre	Instancias	Atributos	Clases
Automobile	205	25	6
Balance	625	4	3
Bupa	345	6	2
Car	1728	6	4
Cleveland	303	13	5
Dermat	366	34	6
German	1000	20	2
Glass	214	9	7
Housevotes	435	16	2
Iris	150	4	3
Mammograph	961	5	2
Pima	768	8	2
Sonar	208	60	2
Spectfheart	267	44	2
Tic-tac-toe	958	9	2
Vehicle	846	18	4
Wisconsin	699	9	2
Zoo	101	16	7
Chess	3196	36	2
Movement-libras	360	90	15
Satimage	6435	36	7
Spambase	4597	57	2
Splice	3190	60	3
Texture	5500	40	11

3. Experimentos

- Hemos evaluado nuestras propuestas mediante una **3x10cv**, y los resultados devueltos son la media los 3 valores devueltos por la media de cada 10cv.
  - Nº generaciones = 50
  - Tamaño población = 20
  - Tamaño población aprendizaje (cada generación) = 10
- **1NN** como algoritmo base de clasificación.

- Atenderemos a los siguientes criterios para comparar nuestras propuestas con otras existentes:
  - Tasa de aciertos (TA)
  - Kappa
  - Tiempo de ejecución
  - % Reducción de atributos

Test de Wilcoxon

3. Experimentos

	IFS-EDAig					IFS-EDAcfs				
Nombre	TA	Kappa	ISR	FSR	Tiempo	TA	Kappa	ISR	FSR	Tiempo
Automobile	65.85	0.55	53.34	69.60	22.21	67.92	0.58	56.49	75.87	1.29
Balance	79.10	0.63	51.59	1.67	5.33	79.53	0.64	53.35	1.67	0.85
Bupa	58.56	0.15	52.93	57.22	5.17	55.58	0.09	53.25	70.56	0.60
Car	90.53	0.80	55.85	16.67	51.04	77.62	0.56	50.32	66.67	2.51
Cleveland	53.33	0.26	54.79	60.51	9.86	54.91	0.28	55.52	54.10	0.98
Dermat	93.17	0.91	54.28	20.59	24.66	94.45	0.93	54.32	59.22	2.48
German	71.70	0.30	51.95	80.17	52.10	68.40	0.24	50.97	80.50	2.43
Glass	67.14	0.55	53.59	35.93	8.31	67.42	0.56	55.36	43.70	0.87
Housevotes	95.49	0.91	52.33	92.71	11.75	95.64	0.91	51.82	93.75	0.51
Iris	94.67	0.92	51.28	68.33	1.10	92.89	0.89	29.58	75.00	0.29
Mammograph	82.13	0.64	51.43	63.33	6.88	80.92	0.62	51.34	42.00	1.49
Pima	70.62	0.35	52.51	54.58	38.65	69.88	0.34	52.66	63.75	1.86
Sonar	74.86	0.49	58.31	68.44	159.43	74.37	0.48	58.23	89.44	5.31
Spectfheart	67.78	0.27	63.24	74.39	109.63	72.90	0.34	58.17	71.29	3.97
Tic-tac-toe	92.63	0.83	52.66	0.00	22.23	69.93	0.34	51.72	88.89	1.16
Vehicle	67.97	0.57	50.25	28.89	74.75	60.68	0.48	52.14	61.85	5.97
Wisconsin	96.37	0.92	50.85	26.67	13.62	95.80	0.91	52.25	32.22	1.53
Zoo	92.03	0.90	54.64	29.38	2.73	90.39	0.87	55.15	55.00	0.31
Ches	95.25	0.90	50.57	58.70	765.57	90.43	0.81	49.89	91.67	15.06
Movement_libras	71.48	0.69	50.61	17.44	1836.38	66.02	0.64	54.02	76.59	40.70
Satimage	89.17	0.87	49.67	15.09	3062.14	88.68	0.86	50.31	36.94	163.46
Spambase	88.70	0.76	50.69	63.98	4969.81	89.07	0.77	50.09	81.35	94.98
Splice	88.45	0.82	48.97	93.11	847.90	88.04	0.81	50.32	90.00	87.89
Texture	98.08	0.98	49.89	15.17	6375.14	96.65	0.96	50.37	74.83	163.43
Media	**81.04**	**0.67**	**52.76**	**46.36**	**769.85**	**78.67**	**0.62**	**51.99**	**65.70**	**25.00**

3. Experimentos

IFS-EDAig

Nombre	TA	Kappa	ISR	FSR	Tiempo
Automobile	65.85	0.55	53.34	69.60	22.21
Balance	79.10	0.63	51.59	1.67	5.33
Bupa	58.56	0.15	52.93	57.22	5.17
Car	90.53	0.80	55.85	16.67	51.04
Cleveland	53.33	0.26	54.79	60.51	9.86
Dermat	93.17	0.91	54.28	20.59	24.66
German	71.70	0.30	51.95	80.17	52.10
Glass	67.14	0.55	53.59	35.93	8.31
Housevotes	95.49	0.91	52.33	92.71	11.75
Iris	94.67	0.92	51.28	68.33	1.10
Mammograph	82.13	0.64	51.43	63.33	6.88
Pima	70.62	0.35	52.51	54.58	38.65
Sonar	74.86	0.49	58.31	68.44	159.43
Spectfheart	67.78	0.27	63.24	74.39	109.63
Tic-tac-toe	92.63	0.83	52.66	0.00	22.23
Vehicle	67.97	0.57	50.25	28.89	74.75
Wisconsin	96.37	0.92	50.85	26.67	13.62
Zoo	92.03	0.90	54.64	29.38	2.73
Ches	95.25	0.90	50.57	58.70	765.57
Movement_libras	71.48	0.69	50.61	17.44	1836.38
Satimage	89.17	0.87	49.67	15.09	3062.14
Spambase	88.70	0.76	50.69	63.98	4969.81
Splice	88.45	0.82	48.97	93.11	847.90
Texture	98.08	0.98	49.89	15.17	6375.14
Media	**81.04**	**0.67**	**52.76**	**46.36**	**769.85**

IFS-EDAcfs

TA	Kappa	ISR	FSR	Tiempo
67.92	0.58	56.49	75.87	1.29
79.53	0.64	53.35	1.67	0.85
55.58	0.09	53.25	70.56	0.60
77.62	0.56	50.32	66.67	2.51
54.91	0.28	55.52	54.10	0.98
94.45	0.93	54.32	59.22	2.48
68.40	0.24	50.97	80.50	2.43
67.42	0.56	55.36	43.70	0.87
95.64	0.91	51.82	93.75	0.51
92.89	0.89	29.58	75.00	0.29
80.92	0.62	51.34	42.00	1.49
69.88	0.34	52.66	63.75	1.86
74.37	0.48	58.23	89.44	5.31
72.90	0.34	58.17	71.29	3.97
69.93	0.34	51.72	88.89	1.16
60.68	0.48	52.14	61.85	5.97
95.80	0.91	52.25	32.22	1.53
90.39	0.87	55.15	55.00	0.31
90.43	0.81	49.89	91.67	15.06
66.02	0.64	54.02	76.59	40.70
88.68	0.86	50.31	36.94	163.46
89.07	0.77	50.09	81.35	94.98
88.04	0.81	50.32	90.00	87.89
96.65	0.96	50.37	74.83	163.43
78.67	**0.62**	**51.99**	**65.70**	**25.00**

3. Experimentos

	IFS-EDAig					IFS-EDAcfs				
Nombre	TA	Kappa	ISR	FSR	Tiempo	TA	Kappa	ISR	FSR	Tiempo
Automobile	65.85	0.55	53.34	69.60	22.21	67.92	0.58	56.49	75.87	1.29
Balance	79.10	0.63	51.59	1.67	5.33	79.53	0.64	53.35	1.67	0.85
Bupa	58.56	0.15	52.93	57.22	5.17	55.58	0.09	53.25	70.56	0.60
Car	90.53	0.80	55.85	16.67	51.04	77.62	0.56	50.32	66.67	2.51
Cleveland	53.33	0.26	54.79	60.51	9.86	54.91	0.28	55.52	54.10	0.98
Dermat	93.17	0.91	54.28	20.59	24.66	94.45	0.93	54.32	59.22	2.48
German	71.70	0.30	51.95	80.17	52.10	68.40	0.24	50.97	80.50	2.43
Glass	67.14	0.55	53.59	35.93	8.31	67.42	0.56	55.36	43.70	0.87
Housevotes	95.49	0.91	52.33	92.71	11.75	95.64	0.91	51.82	93.75	0.51
Iris	94.67	0.92	51.28	68.33	1.10	92.89	0.89	29.58	75.00	0.29
Mammograph	82.13	0.64	51.43	63.33	6.88	80.92	0.62	51.34	42.00	1.49
Pima	70.62	0.35	52.51	54.58	38.65	69.88	0.34	52.66	63.75	1.86
Sonar	74.86	0.49	58.31	68.44	159.43	74.37	0.48	58.23	89.44	5.31
Spectfheart	67.78	0.27	63.24	74.39	109.63	72.90	0.34	58.17	71.29	3.97
Tic-tac-toe	92.63	0.83	52.66	0.00	22.23	69.93	0.34	51.72	88.89	1.16
Vehicle	67.97	0.57	50.25	28.89	74.75	60.68	0.48	52.14	61.85	5.97
Wisconsin	96.37	0.92	50.85	26.67	13.62	95.80	0.91	52.25	32.22	1.53
Zoo	92.03	0.90	54.64	29.38	2.73	90.39	0.87	55.15	55.00	0.31
Ches	95.25	0.90	50.57	58.70	765.57	90.43	0.81	49.89	91.67	15.06
Movement_libras	71.48	0.69	50.61	17.44	1836.38	66.02	0.64	54.02	76.59	40.70
Satimage	89.17	0.87	49.67	15.09	3062.14	88.68	0.86	50.31	36.94	163.46
Spambase	88.70	0.76	50.69	63.98	4969.81	89.07	0.77	50.09	81.35	94.98
Splice	88.45	0.82	48.97	93.11	847.90	88.04	0.81	50.32	90.00	87.89
Texture	98.08	0.98	49.89	15.17	6375.14	96.65	0.96	50.37	74.83	163.43
Media	**81.04**	**0.67**	**52.76**	**46.36**	**769.85**	**78.67**	**0.62**	**51.99**	**65.70**	**25.00**

3. Experimentos

IFS-EDAig

Nombre	TA	Kappa	ISR	FSR	Tiempo
Automobile	65.85	0.55	53.34	69.60	22.21
Balance	79.10	0.63	51.59	1.67	5.33
Bupa	58.56	0.15	52.93	57.22	5.17
Car	90.53	0.80	55.85	16.67	51.04
Cleveland	53.33	0.26	54.79	60.51	9.86
Dermat	93.17	0.91	54.28	20.59	24.66
German	71.70	0.30	51.95	80.17	52.10
Glass	67.14	0.55	53.59	35.93	8.31
Housevotes	95.49	0.91	52.33	92.71	11.75
Iris	94.67	0.92	51.28	68.33	1.10
Mammograph	82.13	0.64	51.43	63.33	6.88
Pima	70.62	0.35	52.51	54.58	38.65
Sonar	74.86	0.49	58.31	68.44	159.43
Spectfheart	67.78	0.27	63.24	74.39	109.63
Tic-tac-toe	92.63	0.83	52.66	0.00	22.23
Vehicle	67.97	0.57	50.25	28.89	74.75
Wisconsin	96.37	0.92	50.85	26.67	13.62
Zoo	92.03	0.90	54.64	29.38	2.73
Ches	95.25	0.90	50.57	58.70	765.57
Movement_libras	71.48	0.69	50.61	17.44	1836.38
Satimage	89.17	0.87	49.67	15.09	3062.14
Spambase	88.70	0.76	50.69	63.98	4969.81
Splice	88.45	0.82	48.97	93.11	847.90
Texture	98.08	0.98	49.89	15.17	6375.14
Media	**81.04**	**0.67**	**52.76**	**46.36**	**769.85**

IFS-EDAcfs

Nombre	TA	Kappa	ISR	FSR	Tiempo
Automobile	67.92	0.58	56.49	75.87	1.29
Balance	79.53	0.64	53.35	1.67	0.85
Bupa	55.58	0.09	53.25	70.56	0.60
Car	77.62	0.56	50.32	66.67	2.51
Cleveland	54.91	0.28	55.52	54.10	0.98
Dermat	94.45	0.93	54.32	59.22	2.48
German	68.40	0.24	50.97	80.50	2.43
Glass	67.42	0.56	55.36	43.70	0.87
Housevotes	95.64	0.91	51.82	93.75	0.51
Iris	92.89	0.89	29.58	75.00	0.29
Mammograph	80.92	0.62	51.34	42.00	1.49
Pima	69.88	0.34	52.66	63.75	1.86
Sonar	74.37	0.48	58.23	89.44	5.31
Spectfheart	72.90	0.34	58.17	71.29	3.97
Tic-tac-toe	69.93	0.34	51.72	88.89	1.16
Vehicle	60.68	0.48	52.14	61.85	5.97
Wisconsin	95.80	0.91	52.25	32.22	1.53
Zoo	90.39	0.87	55.15	55.00	0.31
Ches	90.43	0.81	49.89	91.67	15.06
Movement_libras	66.02	0.64	54.02	76.59	40.70
Satimage	88.68	0.86	50.31	36.94	163.46
Spambase	89.07	0.77	50.09	81.35	94.98
Splice	88.04	0.81	50.32	90.00	87.89
Texture	96.65	0.96	50.37	74.83	163.43
Media	**78.67**	**0.62**	**51.99**	**65.70**	**25.00**

3. Experimentos

IFS-EDAig

Nombre	TA	Kappa	ISR	FSR	Tiempo
Automobile	65.85	0.55	53.34	69.60	22.21
Balance	79.10	0.63	51.59	1.67	5.33
Bupa	58.56	0.15	52.93	57.22	5.17
Car	90.53	0.80	55.85	16.67	51.04
Cleveland	53.33	0.26	54.79	60.51	9.86
Dermat	93.17	0.91	54.28	20.59	24.66
German	71.70	0.30	51.95	80.17	52.10
Glass	67.14	0.55	53.59	35.93	8.31
Housevotes	95.49	0.91	52.33	92.71	11.75
Iris	94.67	0.92	51.28	68.33	1.10
Mammograph	82.13	0.64	51.43	63.33	6.88
Pima	70.62	0.35	52.51	54.58	38.65
Sonar	74.86	0.49	58.31	68.44	159.43
Spectfheart	67.78	0.27	63.24	74.39	109.63
Tic-tac-toe	92.63	0.83	52.66	0.00	22.23
Vehicle	67.97	0.57	50.25	28.89	74.75
Wisconsin	96.37	0.92	50.85	26.67	13.62
Zoo	92.03	0.90	54.64	29.38	2.73
Ches	95.25	0.90	50.57	58.70	765.57
Movement_libras	71.48	0.69	50.61	17.44	1836.38
Satimage	89.17	0.87	49.67	15.09	3062.14
Spambase	88.70	0.76	50.69	63.98	4969.81
Splice	88.45	0.82	48.97	93.11	847.90
Texture	98.08	0.98	49.89	15.17	6375.14
Media	81.04	0.67	52.76	46.36	769.85

IFS-EDAcfs

TA	Kappa	ISR	FSR	Tiempo
67.92	0.58	56.49	75.87	1.29
79.53	0.64	53.35	1.67	0.85
55.58	0.09	53.25	70.56	0.60
77.62	0.56	50.32	66.67	2.51
54.91	0.28	55.52	54.10	0.98
94.45	0.93	54.32	59.22	2.48
68.40	0.24	50.97	80.50	2.43
67.42	0.56	55.36	43.70	0.87
95.64	0.91	51.82	93.75	0.51
92.89	0.89	29.58	75.00	0.29
80.92	0.62	51.34	42.00	1.49
69.88	0.34	52.66	63.75	1.86
74.37	0.48	58.23	89.44	5.31
72.90	0.34	58.17	71.29	3.97
69.93	0.34	51.72	88.89	1.16
60.68	0.48	52.14	61.85	5.97
95.80	0.91	52.25	32.22	1.53
90.39	0.87	55.15	55.00	0.31
90.43	0.81	49.89	91.67	15.06
66.02	0.64	54.02	76.59	40.70
88.68	0.86	50.31	36.94	163.46
89.07	0.77	50.09	81.35	94.98
88.04	0.81	50.32	90.00	87.89
96.65	0.96	50.37	74.83	163.43
78.67	0.62	51.99	65.70	25.00

3. Experimentos

	IFS-EDAig					IFS-EDAcfs				
Nombre	TA	Kappa	ISR	FSR	Tiempo	TA	Kappa	ISR	FSR	Tiempo
Automobile	65.85	0.55	53.34	69.60	22.21	67.92	0.58	56.49	75.87	1.29
Balance	79.10	0.63	51.59	1.67	5.33	79.53	0.64	53.35	1.67	0.85
Bupa	58.56	0.15	52.93	57.22	5.17	55.58	0.09	53.25	70.56	0.60
Car	90.53	0.80	55.85	16.67	51.04	77.62	0.56	50.32	66.67	2.51
Cleveland	53.33	0.26	54.79	60.51	9.86	54.91	0.28	55.52	54.10	0.98
Dermat	93.17	0.91	54.28	20.59	24.66	94.45	0.93	54.32	59.22	2.48
German	71.70	0.30	51.95	80.17	52.10	68.40	0.24	50.97	80.50	2.43
Glass	67.14	0.55	53.59	35.93	8.31	67.42	0.56	55.36	43.70	0.87
Housevotes	95.49	0.91	52.33	92.71	11.75	95.64	0.91	51.82	93.75	0.51
Iris	94.67	0.92	51.28	68.33	1.10	92.89	0.89	29.58	75.00	0.29
Mammograph	82.13	0.64	51.43	63.33	6.88	80.92	0.62	51.34	42.00	1.49
Pima	70.62	0.35	52.51	54.58	38.65	69.88	0.34	52.66	63.75	1.86
Sonar	74.86	0.49	58.31	68.44	159.43	74.37	0.48	58.23	89.44	5.31
Spectfheart	67.78	0.27	63.24	74.39	109.63	72.90	0.34	58.17	71.29	3.97
Tic-tac-toe	92.63	0.83	52.66	0.00	22.23	69.93	0.34	51.72	88.89	1.16
Vehicle	67.97	0.57	50.25	28.89	74.75	60.68	0.48	52.14	61.85	5.97
Wisconsin	96.37	0.92	50.85	26.67	13.62	95.80	0.91	52.25	32.22	1.53
Zoo	92.03	0.90	54.64	29.38	2.73	90.39	0.87	55.15	55.00	0.31
Ches	95.25	0.90	50.57	58.70	765.57	90.43	0.81	49.89	91.67	15.06
Movement_libras	71.48	0.69	50.61	17.44	1836.38	66.02	0.64	54.02	76.59	40.70
Satimage	89.17	0.87	49.67	15.09	3062.14	88.68	0.86	50.31	36.94	163.46
Spambase	88.70	0.76	50.69	63.98	4969.81	89.07	0.77	50.09	81.35	94.98
Splice	88.45	0.82	48.97	93.11	847.90	88.04	0.81	50.32	90.00	87.89
Texture	98.08	0.98	49.89	15.17	6375.14	96.65	0.96	50.37	74.83	163.43
Media	81.04	0.67	52.76	46.36	769.85	78.67	0.62	51.99	65.70	25.00

3. Experimentos

Medida	IFS-EDAig	IFS-EDAcfs	IFS-CoCo	IFS-CHC	IGA	HGA	BDD original
TA	81.0446	78.6722	84.2963	81.3458	71.9675	81.5233	83.1881
Kappa	0.6669	0.6215	0.6963	0.6107	0.6097	0.6482	0.6289
Tiempo	769.8488	24.9978	10356.8192	374.1075	5723.7379	5092.3096	-
ISR	52.7599	51.9850	94.5667	-	-	-	-
FSR	46.3572	65.7023	53.1729	-	-	-	-

3. Experimentos

Medida	IFS-EDAig	IFS-EDAcfs	IFS-CoCo	IFS-CHC	IGA	HGA	BDD original
TA	81.0446	78.6722	84.2963	81.3458	71.9675	81.5233	83.1881
Kappa	0.6669	0.6215	0.6963	0.6107	0.6097	0.6482	0.6289
Tiempo	769.8488	24.9978	10356.8192	374.1075	5723.7379	5092.3096	-
ISR	52.7599	51.9850	94.5667	-	-	-	-
FSR	46.3572	65.7023	53.1729	-	-	-	-

3. Experimentos

Medida	IFS-EDAig	IFS-EDAcfs	IFS-CoCo	IFS-CHC	IGA	HGA	BDD original
TA	81.0446	78.6722	84.2963	81.3458	71.9675	81.5233	83.1881
Kappa	0.6669	0.6215	0.6963	0.6107	0.6097	0.6482	0.6289
Tiempo	769.8488	24.9978	10356.8192	374.1075	5723.7379	5092.3096	-
ISR	52.7599	51.9850	94.5667	-	-	-	-
FSR	46.3572	65.7023	53.1729	-	-	-	-

3. Experimentos

Medida	IFS-EDAig	IFS-EDAcfs	IFS-CoCo	IFS-CHC	IGA	HGA	BDD original
TA	81.0446	78.6722	84.2963	81.3458	71.9675	81.5233	83.1881
Kappa	0.6669	0.6215	0.6963	0.6107	0.6097	0.6482	0.6289
Tiempo	769.8488	24.9978	10356.8192	374.1075	5723.7379	5092.3096	-
ISR	52.7599	51.9850	94.5667	-	-	-	-
FSR	46.3572	65.7023	53.1729	-	-	-	-

3. Experimentos

Medida	IFS-EDAig	IFS-EDAcfs	IFS-CoCo	IFS-CHC	IGA	HGA	BDD original
TA	81.0446	78.6722	84.2963	81.3458	71.9675	81.5233	83.1881
Kappa	0.6669	0.6215	0.6963	0.6107	0.6097	0.6482	0.6289
Tiempo	769.8488	24.9978	10356.8192	374.1075	5723.7379	5092.3096	-
ISR	52.7599	51.9850	94.5667	-	-	-	-
FSR	46.3572	65.7023	53.1729	-	-	-	-

3. Experimentos

Medida	IFS-EDAig	IFS-EDAcfs	IFS-CoCo	IFS-CHC	IGA	HGA	BDD original
TA	81.0446	78.6722	84.2963	81.3458	71.9675	81.5233	83.1881
Kappa	0.6669	0.6215	0.6963	0.6107	0.6097	0.6482	0.6289
Tiempo	769.8488	24.9978	10356.8192	374.1075	5723.7379	5092.3096	-
ISR	52.7599	51.9850	94.5667	-	-	-	-
FSR	46.3572	65.7023	53.1729	-	-	-	-

3. Experimentos

Medida	IFS-EDAig	IFS-EDAcfs	IFS-CoCo	IFS-CHC	IGA	HGA	BDD original
TA	81.0446	78.6722	84.2963	81.3458	71.9675	81.5233	83.1881
Kappa	0.6669	0.6215	0.6963	0.6107	0.6097	0.6482	0.6289
Tiempo	769.8488	24.9978	10356.8192	374.1075	5723.7379	5092.3096	-
ISR	52.7599	51.9850	94.5667	-	-	-	-
FSR	46.3572	65.7023	53.1729	-	-	-	-

3. Experimentos

- IFS-EDA mejor que el resto

Negrita: diferencia estadística

TASA DE ACIERTO

	IFS-CoCo	IFS-CHC	IGA	HGA	BDD original
IFS-EDAig	**0.0010**	0.4387	**0.0003**•	0.4073	**0.0060**
IFS-EDAcfs	**0.0000**	**0.0036**	**0.0035**•	**0.0022**	**0.0060**

KAPPA

	IFS-CoCo	IFS-CHC	IGA	HGA	BDD original
IFS-EDAig	**0.0278**	**0.0108**•	**0.0297**•	0.3216•	0.3115•
IFS-EDAcfs	**0.0003**	0.3658 •	0.4058 •	0.8982•	0.7805

Resultados comparables con IFS-CoCo al tomar sólo bdd discretas

3. Experimentos

- IFS-EDA mejor que el resto

Negrita: diferencia estadística

TASA DE ACIERTO

	IFS-CoCo	IFS-CHC	IGA	HGA	BDD original
IFS-EDAig	**0.0010**	0.4387	**0.0003**●	0.4073	**0.0060**
IFS-EDAcfs	**0.0000**	**0.0036**	**0.0035**●	**0.0022**	**0.0060**

KAPPA

	IFS-CoCo	IFS-CHC	IGA	HGA	BDD original
IFS-EDAig	**0.0278**	**0.0108**●	**0.0297**●	0.3216●	0.3115●
IFS-EDAcfs	**0.0003**	0.3658 ●	0.4058 ●	0.8982●	0.7805

Resultados comparables con IFS-CoCo al tomar sólo bdd discretas

3. Experimentos

- IFS-EDA mejor que el resto

Negrita: diferencia estadística

TASA DE ACIERTO

	IFS-CoCo	IFS-CHC	IGA	HGA	BDD original
IFS-EDAig	**0.0010**	0.4387	**0.0003**•	0.4073	**0.0060**
IFS-EDAcfs	**0.0000**	**0.0036**	**0.0035**•	**0.0022**	**0.0060**

KAPPA

	IFS-CoCo	IFS-CHC	IGA	HGA	BDD original
IFS-EDAig	**0.0278**	**0.0108**•	**0.0297**•	0.3216•	0.3115•
IFS-EDAcfs	**0.0003**	0.3658 •	0.4058 •	0.8982•	0.7805

Resultados comparables con IFS-CoCo al tomar sólo bdd discretas

4. Conclusiones

- Propuesta de dos algoritmos basados en EDAs para la IFS:
  - **IFS-EDAig**: UMDA para un nº creciente de atributos evaluados con IG. Validación cruzada de 1NN para seleccionar nº atributos final en función de instancias.
  - **IFS-EDAcfs:** más sencillo. Un único UMDA con selección de atributos mediante métrica cfs.
- **Ventajas** a nivel de kappa con el resto de técnicas excepto IFS-CoCo, a quien aventaja en coste computacional.
- Otras ventajas:
  - Necesidad de ajustar un nº menor de parámetros .
  - Posibilidad de recurrir a la paralelización.
  - Las bases de datos resultantes están balanceadas.
  - Porcentaje de reducción en cuanto al número de atributos es mayor.

Trabajo futuro

- Hacer extensibles los algoritmos a bases de datos con atributos numéricos para evitar el uso de técnicas de discretización .
  - Medidas filter que manejen los **valores numéricos directamente**.
- Explotar formas alternativas de inicializar la población para **aumentar el porcentaje de reducción de instancias**.
- Probar diferentes **EDAs bivariados**, o incluso multivariados, para manejar las relaciones entre las variables.
- Interesante: adaptar las propuestas HGA e IFS-CoCo para la utilización de EDAs.

Gracias!