

Métodos para Determinar el Atributo Distinguido en Multiredes Bayesianas

Andrés Cano Utrera
Fco. Javier García Castellano
Andrés R. Masegosa Arredondo
Serafín Moral Callejón

Uncertainty Treatment in Artificial Intelligence Research Group
Department of Computer Science and Artificial Intelligence

Granada University (Spain)



Introducción



Objetivo: Utilización en Elvira de multiredes bayesianas para clasificación.



Estado:



Cualquier clasificador (interfaz Classifier) puede ser utilizado en las multiredes.



Multiredes con atributo predictivo como variable distinguida.



Multiredes con variables a clasificar como variable distinguida (por hacer).



Multiredes recursivas (incluir más criterios de parada).

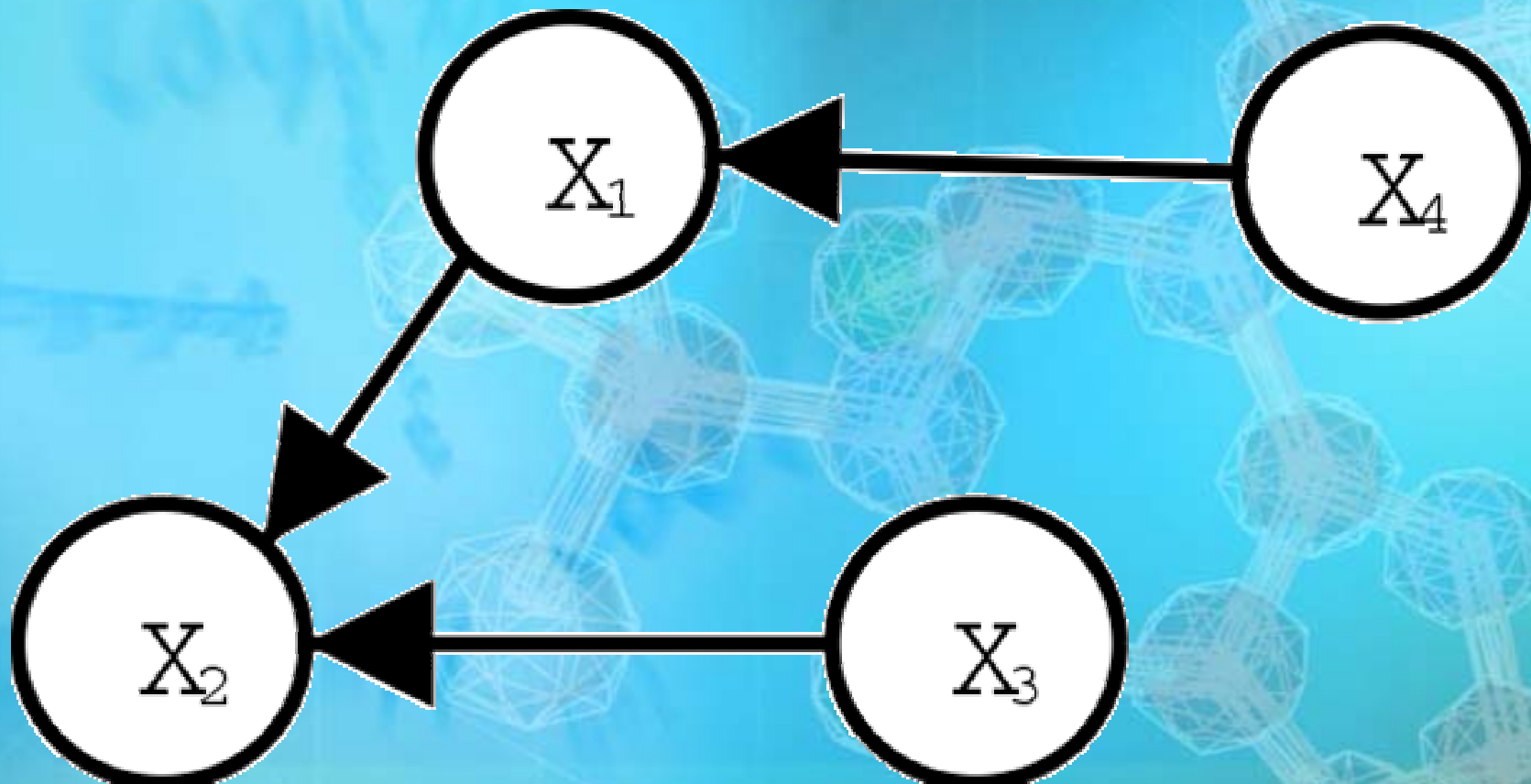


No utilizables desde el interfaz gráfica.

Introducción

Multiredes Bayesianas

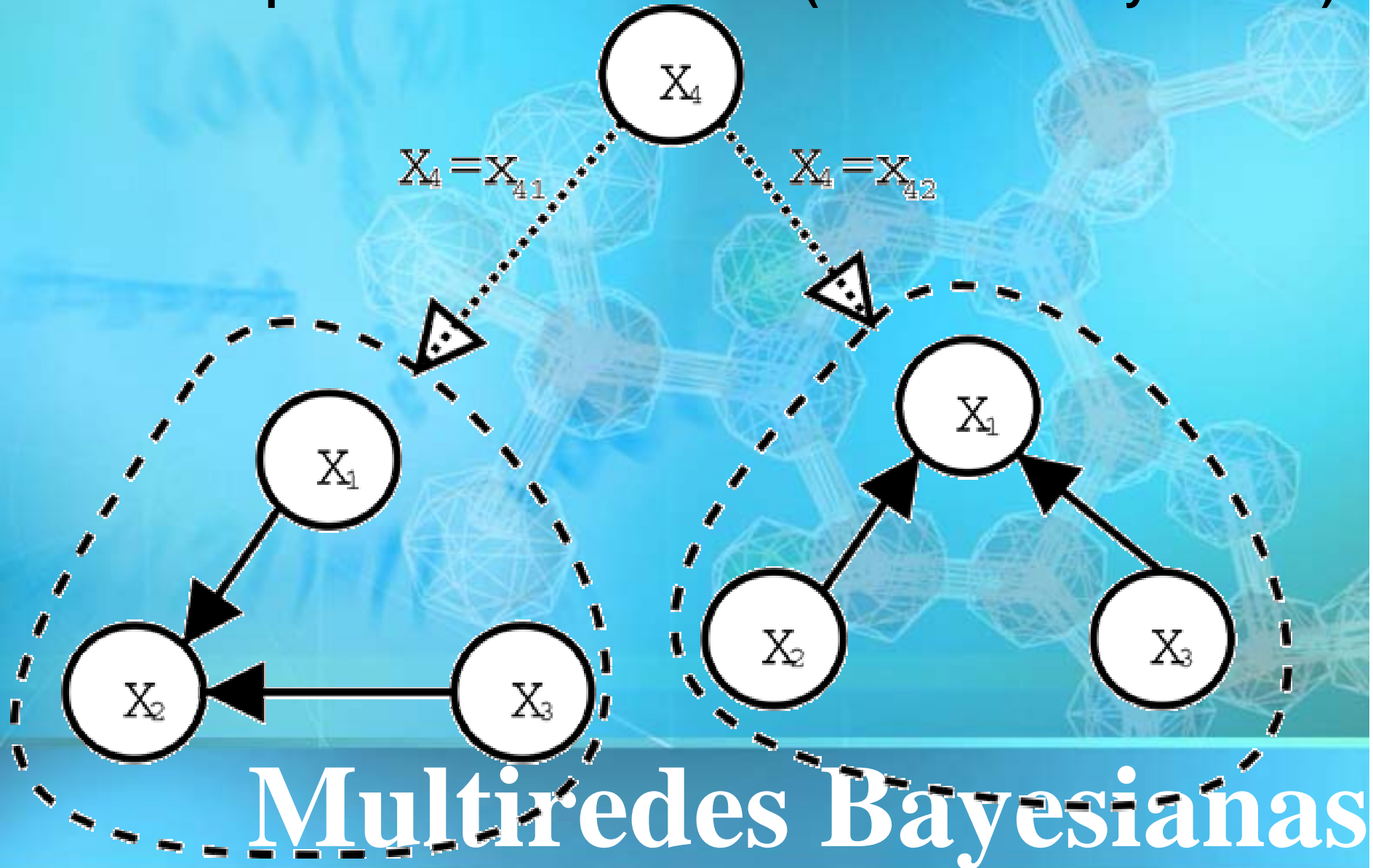
- (In)dependencias no específicas del contexto o (in)dependencias simétricas (Redes Bayesianas)



Multiredes Bayesianas

Multiredes Bayesianas

- (In)dependencias específicas del contexto o independencias asimétricas (Multiredes Bayesianas).



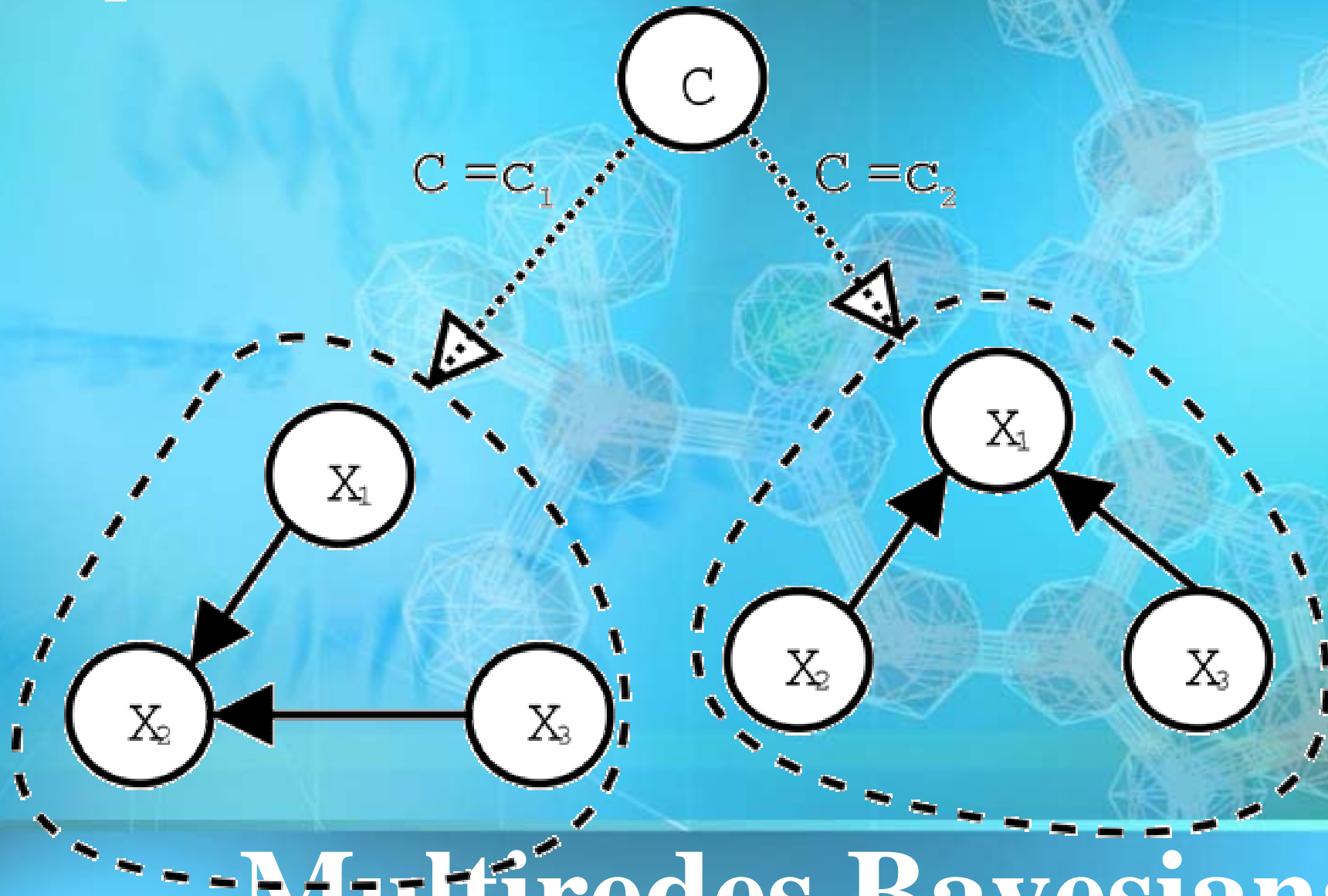
Multiredes Bayesianas en clasificación

- 🕒 **Las redes bayesianas en las multiredes pueden ser clasificadores bayesianos**
- 🕒 **En clasificación Heckerman distingue entre:**
 - 🌐 **Independencias asimétricas de subconjunto:** Distintas redes bayesianas para distintos valores de la variable a clasificar.
 - 🌐 **Independencias asimétricas de hipótesis específica:** Distintas redes bayesianas para distintos valores una variable predictora.

Multiredes Bayesianas

Multiredes Bayesianas en clasificación

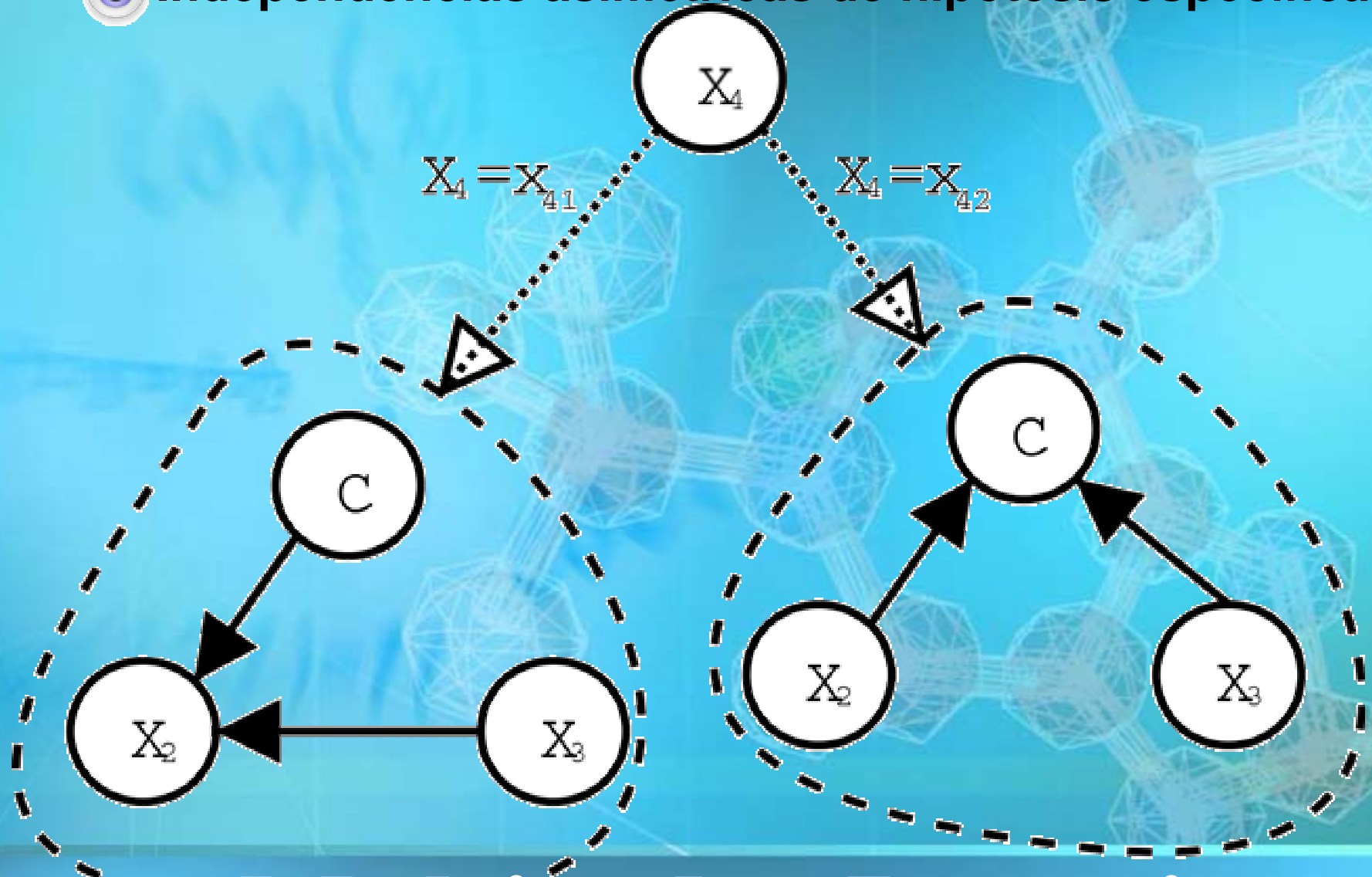
Independencias asimétricas de subconjunto



Multiredes Bayesianas

Multiredes Bayesianas en clasificación

Independencias asimétricas de hipótesis específica



Multiredes Bayesianas

Multiredes Bayesianas en clasificación



Multiredes Bayesianas recursivas. Criterios de Parada.



Variable distinguida. Aproximaciones:



Enfoque Wrapper: Se construye una multired para cada posible variable distinguida, se valida cada multired. Accuracy y Log Likelihood.



Enfoque Filter: Se ordenan las variables según una función.



Enfoque Híbrido: Se ordenan las variables según una función, a un subconjunto de las mejores variables se le aplica una búsqueda wrapper.

Multiredes Bayesianas

Funciones Filter en Multiredes Bayesianas

- KullbackLeibler1

- KullbackLeibler2

- Bhattacharyya

- Matusita

- entropyShanon

- Gain

- Gain ratio

- Gain Dirichlet

- Score attribute as parent of the rest

- Score attribute parent of Class

- Conditional Mutual Information

- Euclidean Distance

- Heuristic $\text{sum}(\text{score}(X \rightarrow C \rightarrow Y) - \text{score}(X \rightarrow C \rightarrow Y \leftarrow X))$

- Heuristic Conditional Mutual Information Max X Of $\text{sum}_i \text{MI}(Y_i, X | C)$

Funciones Filter

Experimentos

- **Multired wrapper (BMN_w) usando tanto por ciento de bien clasificados (accuracy) 5-CV**
- **14 multiredes filter (BMN_f)**
- **24 problemas de *UCI repository of machine learning databases***
- **Valores perdidos eliminados**
- **Discretización basada en la entropía (Fayyad & Irani)**
- **Validación cruzada 10-veces**
- **Test de signos pareados de Wilcoxon**
- **Naïve-Bayes en las hojas**

Experimentos

Experimentos: Multiredes Filter

- **Filter distancia de Kullback-Leibler (BMN_{f_1}):**
 - Método para medir la distancia entre dos distribuciones de probabilidad.
 - Se mide el grado de dependencia para cada atributo y la clase.
 - La idea es escoger aquella variable con un grado más alto de dependencia con la variable a clasificar.

$$D_{kl} (P(X), Q(X)) = \sum_{x_i} P(x_i) \log \frac{P(x_i)}{Q(x_i)}$$

$$KL_{ij}(X; C)_2 = D_{kl} (P(X|c_i), P(X|c_j)) + D_{kl} (P(X|c_j), P(X|c_i))$$

Experimentos: Multiredes Filter

- **Filter distancia de Matusita (BMN_{f2}):**
 - Mide la distancia entre dos distribuciones de probabilidad.
 - Se mide el grado de dependencia para cada atributo y la clase, considerando la distancia media de las distribuciones condicionales.

$$D_m(X; C) = \sum_{i=1}^{r_c} \sum_{j=1}^{j < i} P(c_i)P(c_j) \left[\sum_{k=1}^{r_x} \left(1 - \sqrt{P(x_k|c_i)P(x_k|c_j)} \right) \right]$$

Experimentos: Multiredes Filter

- **Filter Ganancia (BMN_{f3}):**
 - Usado por Quinlan en ID3.
 - También conocido como Información Mutua.
 - Método más usado para medir el grado de dependencia de un atributo con la clase.

$$Gain(X; C) = H(C) - H(C|X) = \sum_{i=1}^{r_x} \sum_{j=1}^{r_c} P(x_i, c_j) \log \frac{P(x_i, c_j)}{P(x_i)P(c_j)}$$

Experimentos: Multiredes Filter

- **Filter razón de Ganancia (BMN_{f4}):**
 - Usado por Quinlan en C4.5
 - Es una modificación de la ganancia para evitar el sesgo que tiene la ganancia en favor de variables con un gran número de casos.

$$GainRatio(X; C) = \frac{Gain(X; C)}{SplitInfo(X)}$$

$$SplitInfo(X) = \sum_{j=1}^{r_x} p(x_j) \log_2(1/p(x_j))$$

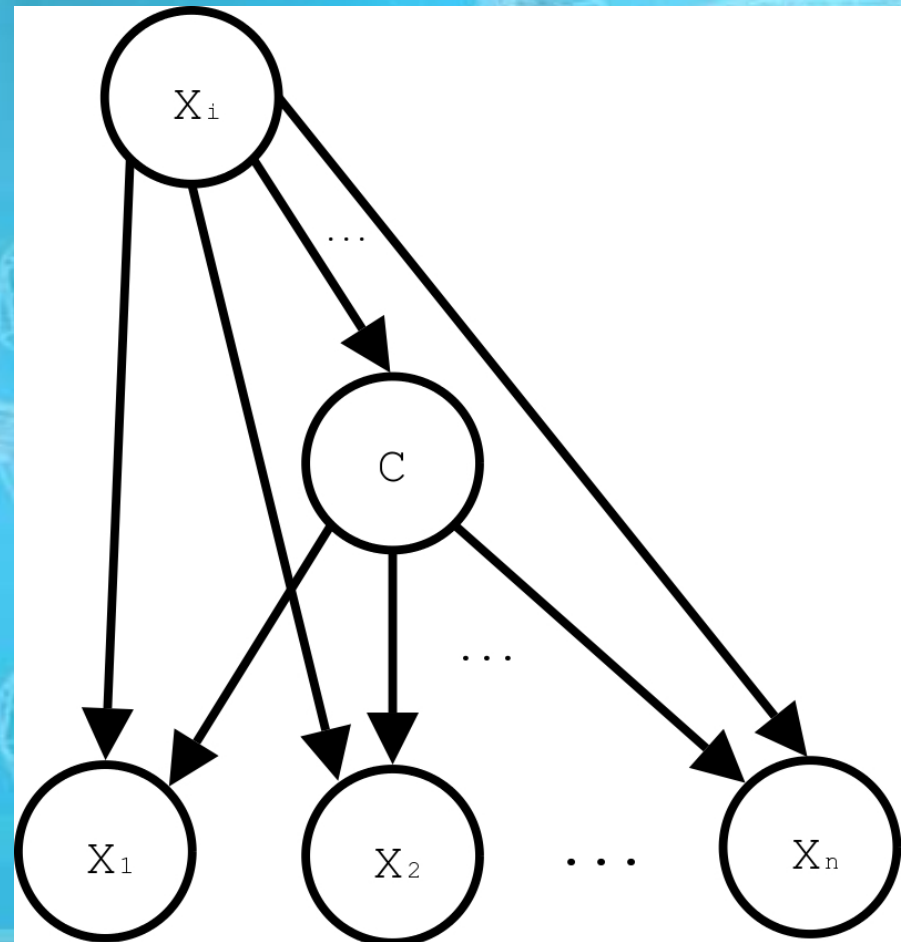
Experimentos: Multiredes Filter

- Filter heurística métrica de un atributo como padre del resto (BMN_{f5} , BMN_{f6} , BMN_{f7}):

- Se construye una red bayesiana para cada atributo X_i , donde dicho atributo es padre del resto de las variables incluida la clase. El resto de las variables predictoras forman una estructura Naïve Bayes con la clase.

- Se calcula la métrica de la red bayesiana construida, para ellos usamos tres métricas K2 (BMN_{f5}), BIC (BMN_{f6}) y BDe (BMN_{f7}).

- Esta estructura es equivalente a una multired de Naïve-Bayes.



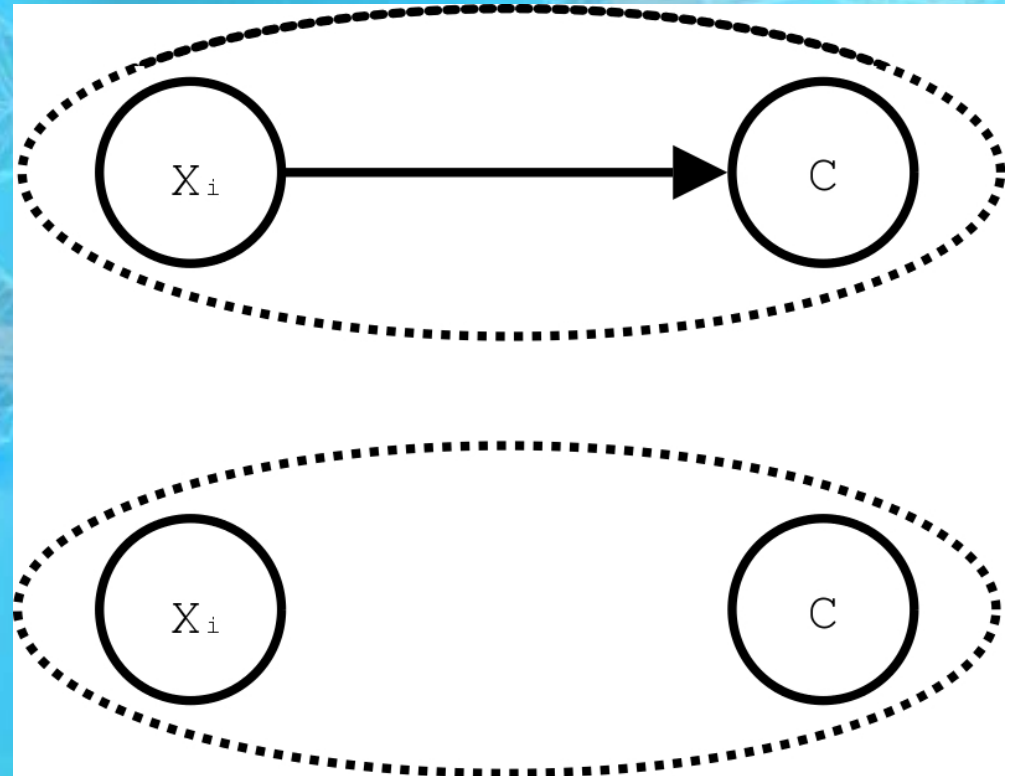
Experimentos

Experimentos: Multiredes Filter

- Filter heurística métrica de un atributo como padre de la clase (BMN_{f8} , BMN_{f9} , BMN_{f10}):

- Se construye dos redes bayesianas para cada atributo X_i , en una, dicho atributo es padre del resto de la clase, en la otra red no. El resto de las variables predictoras no se tienen en cuenta.

- Se calcula la métrica para las dos redes y se calcula la diferencia, para ello usamos tres métricas K2 (BMN_{f8}), BIC (BMN_{f9}) y BDe (BMN_{f10}).



Experimentos: Multiredes Filter

- **Filter información mutua condicionada (BMN_{f11}):**
 - Para cada variable X_i se obtiene la sumatoria de la información mutua condicional dada la clase con la variable X_j ($i \neq j$)
 - Esta medida tiene en cuenta la relación de cada variable X_i con el resto de las variables

$$CMufInf_{f11}(X_i) = \sum_j MI(X_i, X_j|C).$$

Experimentos: Multiredes Filter

- **Filter Bhattacharyya (BMN_{f12}):**
 - Mide la distancia entre dos distribuciones de probabilidad.
 - Se mide el grado de dependencia para cada atributo y la clase, comportamiento muy parecido a la distancia de Matusita.

$$Bh(X; C) = \sum_{i=1}^{r_c} -\log \left[P(c_i) \sum_{j=1}^{r_x} \sqrt{P(x_j|c_i)P(x_j)} \right]$$

Experimentos: Multiredes Filter

- **Filter Ganancia suavizada (BMN_{f13} , BMN_{f14}):**
 - Usada en árboles de clasificación
 - Se trata de una ganancia suavizada para evitar sobreajuste a los datos.
 - Se tiene dos versiones ($BMNf13$, $BMNf14$), según el suavizado de los datos ($s=1$, $s=2$).

$$P(c_j) = freq(c_j)/n$$

$$p^d(c_j) = (s/r_c + freq(c_j))/(s + n)$$

Experimentos: Problemas

#	Problema	Casos	Atributos	Clases	#	Problema	Casos	Atributos	Clases
1	australian	690	14	2	14	iris	150	4	3
2	breast	682	10	2	15	letter	20000	16	26
3	chess	3196	36	2	16	lymphography	148	18	4
4	cleve	296	13	2	17	mofn-3-7-10	1324	10	2
5	corral	128	6	2	18	pima	768	8	2
6	crx	653	15	2	19	satimage	6435	36	6
7	diabetes	768	8	2	20	segment	2310	19	7
8	flare	1066	10	2	21	shuttle-small	5800	9	7
9	german	1000	20	2	22	soybean-large	562	35	19
10	glass	214	9	7	23	vehicle	846	18	4
11	glass2	163	9	2	24	vote	435	16	2
12	heart	270	13	2	25	waveform-21	5000	21	3
13	hepatitis	80	19	2					

Experimentos

Experimentos: Resultados (1)

#	w	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14
1	85,80	86,38	86,38	86,38	86,38	86,52	86,52	86,52	86,38	86,38	86,38	84,49	86,38	86,38	86,38
2	95,90	96,19	96,19	96,19	95,31	96,19	96,19	96,62	96,19	96,19	96,19	96,19	96,19	96,19	96,19
3	94,27	93,52	93,52	93,52	93,52	84,58	84,58	84,58	93,52	93,52	93,52	84,58	93,52	93,52	93,52
4	81,41	81,78	82,12	82,12	82,09	82,44	82,78	82,78	81,78	81,78	81,78	83,82	82,12	81,78	81,78
5	89,10	85,19	85,19	85,19	85,19	87,56	88,40	87,56	85,19	85,19	85,19	88,40	85,19	85,19	85,19
6	85,74	86,97	86,97	86,97	86,97	86,97	86,66	86,97	86,97	86,97	86,97	85,75	86,97	86,97	86,97
7	76,95	77,07	77,07	77,07	77,07	77,61	77,61	77,61	77,07	77,07	77,07	77,61	77,07	77,07	77,07
8	82,08	81,80	81,80	81,80	78,89	81,80	81,80	81,80	81,80	81,80	81,80	81,80	81,80	81,80	81,80
9	73,90	75,20	75,20	75,20	75,20	73,00	74,90	75,50	75,20	75,20	75,20	73,50	75,20	75,20	75,20
10	74,85	71,06	72,97	72,99	72,47	75,78	73,42	73,42	74,85	74,85	74,85	72,97	72,97	73,46	72,99
11	84,71	84,71	84,71	84,71	82,87	84,71	84,12	85,33	84,71	84,71	84,71	84,12	84,71	84,71	84,71
12	81,48	82,96	82,96	82,96	82,96	82,22	83,70	83,70	82,96	82,96	82,96	83,33	82,96	82,96	82,96
13	87,50	87,50	87,50	87,50	87,50	90,00	85,00	85,00	87,50	87,50	87,50	88,75	87,50	87,50	88,75
14	93,33	94,67	94,00	94,00	94,00	94,00	94,67	93,33	94,00	94,00	94,00	94,67	94,00	94,00	94,00
15	83,57	82,40	82,40	82,40	78,93	83,57	73,99	79,91	82,40	82,40	82,40	83,57	82,40	82,40	82,40
16	82,43	82,33	79,67	79,67	81,05	82,43	85,10	84,48	82,33	82,33	82,33	77,81	83,05	82,33	82,33
17	93,81	93,88	93,88	94,03	94,03	94,79	94,79	94,79	94,03	94,03	94,03	94,71	93,88	94,03	94,03
18	77,73	77,47	77,47	77,47	77,47	77,73	77,73	77,73	77,47	77,47	77,47	77,73	77,47	77,47	77,47
19	86,01	84,29	85,21	84,77	84,29	86,01	85,49	86,01	84,29	85,11	84,29	86,01	84,29	84,77	84,77
20	94,98	93,81	93,81	93,81	93,20	95,02	92,08	95,02	93,81	93,81	93,81	95,02	93,81	93,81	93,81
21	99,60	99,17	99,17	99,17	99,17	99,53	99,00	99,53	99,17	99,17	99,17	99,17	99,17	99,17	99,17
22	91,81	91,64	85,58	91,46	91,46	91,81	91,46	92,00	91,64	91,64	91,64	85,58	85,58	91,64	91,64
23	69,76	71,17	71,17	71,17	67,39	70,82	69,88	70,47	71,17	70,70	70,23	71,17	71,17	71,17	71,17
24	94,28	92,20	92,20	92,20	92,20	93,12	93,12	93,12	92,20	92,20	92,20	93,82	92,20	92,20	92,20
25	82,94	81,72	81,72	81,72	81,52	81,52	81,52	81,52	81,72	81,72	81,72	81,72	81,72	81,72	81,72
t	1	140	130	80	80	7	21	7	65	99	65	4	142	79	79

Experimentos

Experimentos: Resultados (2)





	w	$f1$	$f2$	$f3$	$f4$	$f5$	$f6$	$f7$	$f8$	$f9$	$f10$	$f11$	$f12$	$f13$	$f14$
w	—	13/7	13/6	13/5	16/8	8/2	13/6	10/3	12/6	12/5	12/6	13/4	12/6	13/5	13/7
$f1$	10/0	—	3/1	3/0	10/3	6/1	11/3	6/2	1/0	2/0	2/1	7/2	2/1	1/0	1/0
$f2$	10/0	3/1	—	1/0	10/4	5/1	10/3	6/2	2/1	3/0	3/2	6/1	1/1	2/0	2/0
$f3$	10/0	4/0	3/1	—	10/3	5/1	10/3	6/2	2/0	2/0	3/1	8/2	4/1	1/0	1/0
$f4$	8/0	3/0	3/1	1/0	—	3/1	7/2	3/1	1/0	1/0	1/0	6/2	2/1	1/0	1/0
$f5$	13/0	16/4	16/4	16/3	20/5	—	11/4	5/2	16/4	17/3	17/4	9/3	15/6	16/3	16/3
$f6$	11/0	11/0	13/1	13/0	17/1	6/1	—	3/0	11/0	11/0	11/0	8/2	13/2	11/0	12/0
$f7$	13/0	18/3	18/3	18/2	21/5	7/1	11/3	—	17/3	17/2	18/3	11/4	18/5	17/2	18/2
$f8$	10/0	2/0	4/1	3/0	10/2	5/1	12/3	7/2	—	1/0	1/1	8/2	3/1	1/0	1/0
$f9$	10/0	3/0	4/1	4/0	11/2	4/1	12/3	7/2	1/0	—	2/1	8/2	4/2	2/0	2/0
$f10$	10/0	2/0	4/1	3/0	10/1	4/1	12/3	6/2	0/0	0/0	—	8/2	3/1	1/0	1/0
$f11$	10/0	12/3	12/2	12/2	18/4	8/0	9/2	8/1	12/3	13/2	13/3	—	12/3	12/2	11/2
$f12$	11/0	3/0	1/0	1/0	10/2	6/1	10/3	6/2	2/0	3/0	3/0	6/1	—	2/0	2/0
$f13$	10/0	3/0	4/1	3/0	11/2	5/1	12/3	7/2	1/0	1/0	2/0	8/2	4/1	—	1/0
$f14$	11/1	4/0	5/1	3/0	12/2	5/1	11/3	6/2	2/0	2/0	3/0	8/2	5/1	1/0	—

Experimentos

Experimentos: Resultados (3)

- Se obtienen mejores resultados en el tanto por ciento de bien clasificados con la multired wrapper (búsqueda exhaustiva)
- La multired wrapper va mejor con bases de datos con muchas variables (Chess, Waveform-21, Letter, Satimage)
- El método wrapper es muy lento
- El método wrapper también es mejor resultados significativos estadísticamente
- BMN_{f_5} (heurística un atributo como padre del resto usando métrica K2) funciona de forma muy aproximada a la wrapper. Similares: BMN_{f_7} , BMN_{f_9} , $BMN_{f_{11}}$
- $BMN_{f_{12}}$ (función Bhattacharyya) funciona muy bien tardando muy poco tiempo.

Conclusiones

-  Se han realizado un estudio sobre distintas multiredes bayesianas donde una variable predictora es la variable por la cual se ramifica.
-  Se ha estudiado la elección de la variable predictora mediante un método wrapper y varios filter.
-  Dependiendo de nuestras necesidades hay distintas alternativas (BMNw, BMNf5, BMNf12)
-  Estamos realizando estudios: utilizando otros clasificadores bayesianos en las hojas, usando metodologías híbridas y multiredes recursivas.

Conclusiones