

Konputazio Zientziak eta Adimen Artifiziala Saila
Departamento de Ciencias de la Computación e Inteligencia Artificial



Euskal Herriko Unibertsitatea
Universidad del País Vasco

**Advances in Supervised Classification
based on Probabilistic Graphical Models**

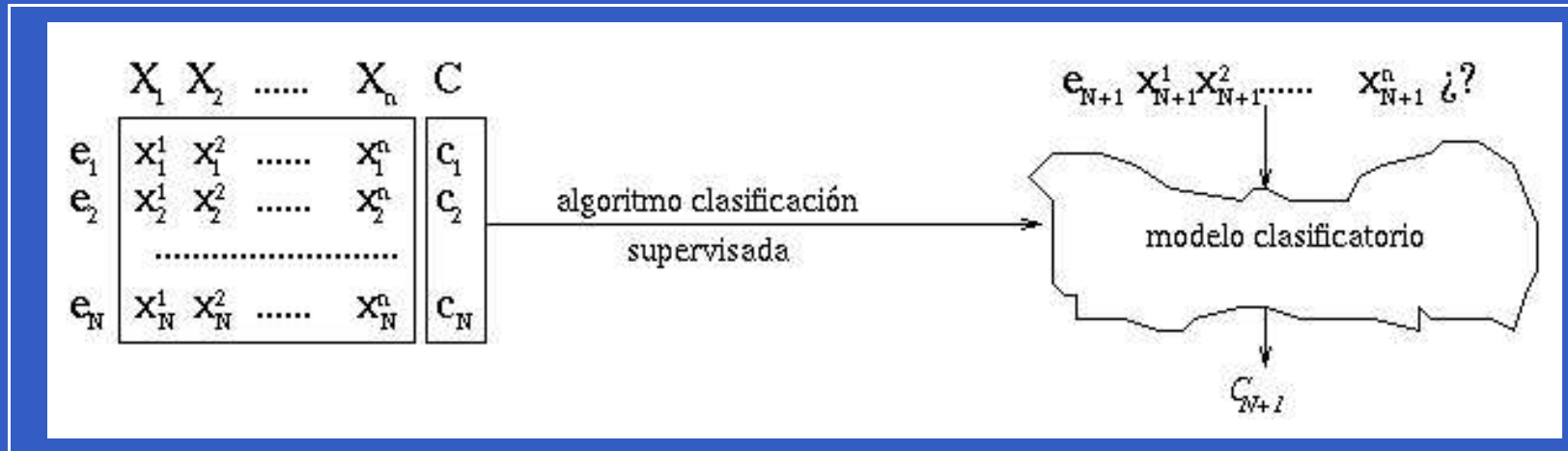
Iñaki Inza

Donostia - San Sebastián, Ekaina - Junio 2002

Tabla de contenidos

- 1: Clasificación Supervisada
- 2: Modelos Gráficos Probabilísticos
- 3: Algoritmos de Estimación de Distribuciones
- 4: Aportación 1: Representación del comportamiento conjunto de clasificadores supervisados
- 5: Aportación 2: Selección de variables para clasificadores supervisados
- 6: Aportación 3: Pesado de atributos para el clasificador del vecino más próximo
- 7: Conclusiones
- 8: Publicaciones

Clasificación supervisada: proceso



- Aspectos a tener en cuenta:
 - Bondad predictiva del clasificador.
 - Coste computacional.
 - Comprensibilidad e interpretabilidad del clasificador.
 - Simplicidad del clasificador: KISS, Occam's Razor.

Clasificación supervisada: historia y familias

- Clasificadores inspirados en la Estadística:
 - Análisis Discriminante.
 - Regresión Logística.
 - ...
- Clasificadores de segunda generación: Aprendizaje Automático ('Machine Learning'):
 - Árboles de clasificación: CART, C4.5, C5.0, ID3,...
 - Reglas de decisión IF-THEN: AQ, CN2, Ripper,...
 - Vecino más próximo: IB, Protos, TiMBL,...
 - ...

Clasificación supervisada: dominios y validación

Dominio	Número de instancias	Número de variables	Número de clases
<i>Echocardiogram</i>	131	7	2
<i>Hepatitis</i>	155	19	2
<i>Glass</i>	214	9	7
<i>Audiology</i>	226	69	24
<i>Heart disease</i>	270	13	2
<i>Breast cancer</i>	286	9	2
<i>Cleveland</i>	303	13	2
<i>Liver (BUPA)</i>	345	6	2
<i>Ionosphere</i>	351	34	2
<i>Arrhythmia</i>	452	279	16
<i>CRX</i>	690	15	2
<i>Diabetes (Pima)</i>	768	8	2
<i>Anneal</i>	898	38	6
<i>Cloud</i>	1,834	204	4
<i>Sick-euthyroid</i>	3,163	25	2
<i>Hypothyroid</i>	3,163	25	2
<i>DNA</i>	3,186	180	3
<i>Int. adv.</i>	3,279	1,558	2
<i>Spambase</i>	4,601	57	2

- Estimación de la bondad predictiva: 10-fold cross-validation (*t*-test apareado),
5 x 2-fold cross-validation (5x2cv *F*-test, Alpaydin'99).

Modelos gráficos probabilísticos: definición

$\mathbf{X} = (X_1, X_2, \dots, X_n)$ Modelo Gráfico Probabilístico (MGP) para \mathbf{X} : factorización gráfica de su distribución de probabilidad generalizada conjunta

2 componentes, MGP = (S, θ_S) :

- Estructura S :
 - Un nodo \longleftrightarrow Una variable.
 - Arco entre dos nodos.
 - S representa explícitamente las (in)dependencias condicionales en \mathbf{X} :

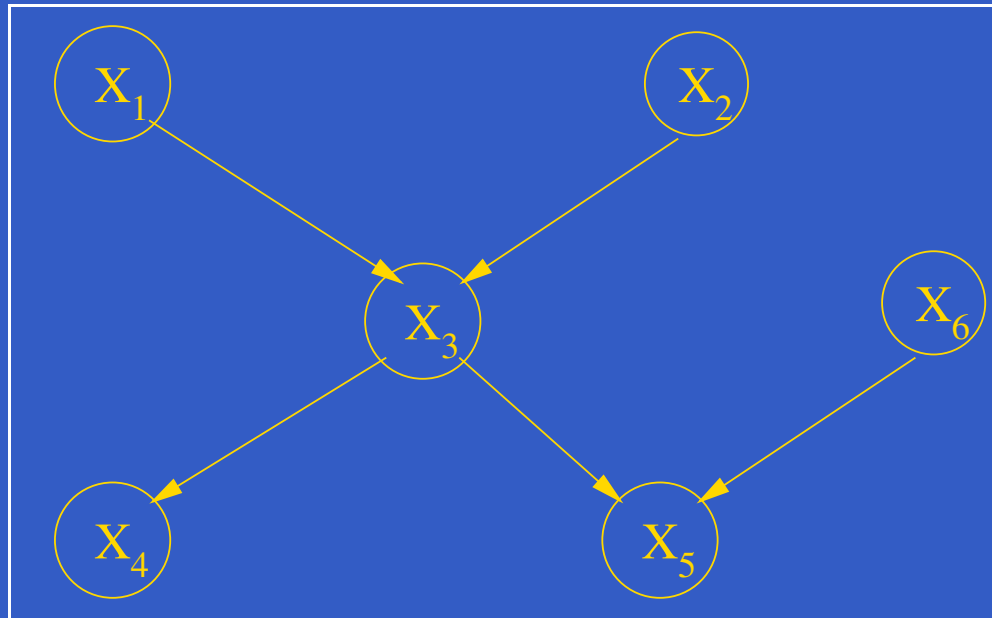
$$I(\mathbf{Y}, \mathbf{Z} \mid \mathbf{W}) \leftrightarrow \rho(\mathbf{y} \mid \mathbf{z}, \mathbf{w}) = \rho(\mathbf{y} \mid \mathbf{w}).$$

- Una variable es condicionalmente independiente de sus no-descendientes, dados sus padres en S (\mathbf{Pa}_i^S):

$$\rho(\mathbf{x}) = \prod_{i=1}^n \rho(x_i \mid \mathbf{pa}_i^S).$$

- Parámetros θ_S : conjunto de distribuciones locales de probabilidad generalizadas, marginales y condicionadas.

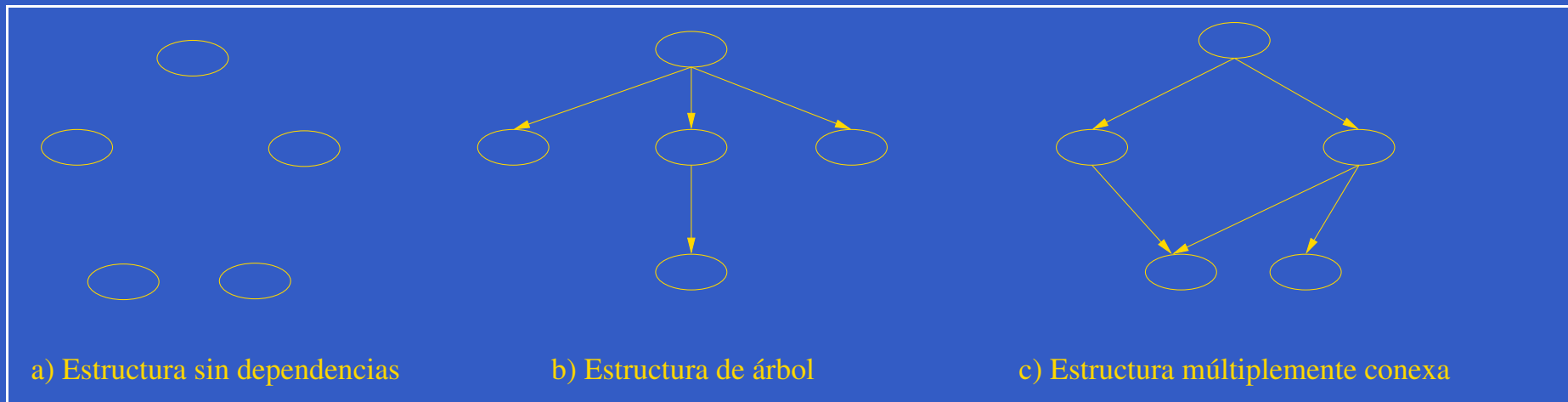
Modelos gráficos probabilísticos: ejemplo



Estructura para un MGP definido sobre $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5, X_6)$.

$$\rho(x_1, x_2, x_3, x_4, x_5, x_6) = \\ \rho(x_1) \cdot \rho(x_2) \cdot \rho(x_3|x_1, x_2) \cdot \rho(x_4|x_3) \cdot \rho(x_5|x_3, x_6) \cdot \rho(x_6).$$

Modelos gráficos probabilísticos: grados de complejidad



- Estructuras sin dependencias: $\rho(\mathbf{x} \mid \boldsymbol{\theta}_S) = \prod_{i=1}^n \rho(x_i \mid \boldsymbol{\theta}_i)$.
- Estructuras con relaciones bivariadas:
 $\rho(\mathbf{x} \mid \boldsymbol{\theta}_S) = \prod_{i=1}^n \rho(x_i \mid x_{j(i)}, \boldsymbol{\theta}_i)$.
- Estructuras múltiplemente conexas:
 $\rho(\mathbf{x} \mid \boldsymbol{\theta}_S) = \prod_{i=1}^n \rho(x_i \mid \text{pa}_i^S, \boldsymbol{\theta}_i)$.

Modelos gráficos probabilísticos: redes Bayesianas

- Variables discretas. Grafo acíclico dirigido.

$$p(x_i^k \mid \mathbf{pa}_i^{j,S}, \theta_i) = \theta_{x_i^k \mid \mathbf{pa}_i^j} \equiv \theta_{ijk}$$

- Problemas NP-duros: propagación de la evidencia (Lauritzen y Spiegelhalter'88) e inducción del modelo (Chickering y col.'94).

Inducción del modelo: estructura y parámetros de probabilidad.

- Proporcionado por un experto.
- Proceso automático a partir de un fichero de casos:
 - Detectando (in)dependencias condicionales en la estructura.
 - Problema de optimización: 'score + search'.
 - Evaluación ('score') de cada estructura: penalización de la verosimilitud (AIC, BIC), métricas Bayesianas (BDe, K2), métricas basadas en la Teoría de la Información (KL, MI, MDL).
 - Búsqueda ('search') en el espacio de estructuras: secuenciales, búsqueda tabú, algoritmos genéticos,...

Modelos gráficos probabilísticos: redes Gaussianas

- Variables continuas. Grafo acíclico dirigido.

$$f(x_i \mid \mathbf{pa}_i^S, \theta_i) \equiv \mathcal{N}(x_i; m_i + \sum_{x_j \in \mathbf{pa}_i} b_{ji}(x_j - m_j), v_i)$$

- Inducción del modelo: estructura y funciones de densidad.
 - Detectando (in)dependencias condicionales en la estructura.
 - Problema de optimización: ‘score + search’.
 - Evaluación (‘score’) de cada estructura: penalización de la verosimilitud, métricas Bayesianas (BGe).
 - Búsqueda (‘search’) en el espacio de estructuras: secuenciales.

Algoritmos de estimación de distribuciones: motivación

EDAs \equiv 'Estimation of Distribution Algorithms'

- Críticas a los algoritmos genéticos (GAs \equiv 'Genetic Algorithms'):
 - Su gran número de parámetros y su proceso de 'tuning'.
 - La difícil predicción de los 'movimientos' de la población en el espacio de búsqueda.
 - Su incapacidad de resolver los problemas 'deceptivos-engañosos'.
 - 'Ruptura' de relaciones relevantes:

* * * 1 0 * * 1 * * *	* * * ? ? * * ? * * *
* * * * * * * * * * *	* * * ? ? * * ? * * *

Algoritmos de estimación de distribuciones: motivación

- De esta manera . . . el descubrimiento de áreas del espacio con valores favorables de la función objetivo puede demorarse.
- GAs: *implícitamente* capturan las relaciones entre las variables. No guardan una información *explícita* acerca de los grupos de variables relacionadas.
- Es importante: *identificación de los grupos de variables relacionadas*, que conjuntamente producen soluciones favorables.
- Primera aproximación: messy Genetic Algorithms (mGA, Goldberg y col.'89).
- En vez de seguir extendiendo el paradigma de búsqueda genético → EDAs (Mühlenbein y Paaß'96).

Algoritmos de estimación de distribuciones: orígenes

EDA

$D_0 \leftarrow$ Generar M individuos (la población inicial) al azar

Repetir para $l = 1, 2, \dots$ hasta la condición de parada

$D_{l-1}^{Se} \leftarrow$ *Seleccionar* $N \leq M$ individuos de D_{l-1} siguiendo un método de selección

$p_l(\mathbf{x}) = p(\mathbf{x} | D_{l-1}^{Se}) \leftarrow$ *Estimar* la distribución de probabilidad de los individuos seleccionados

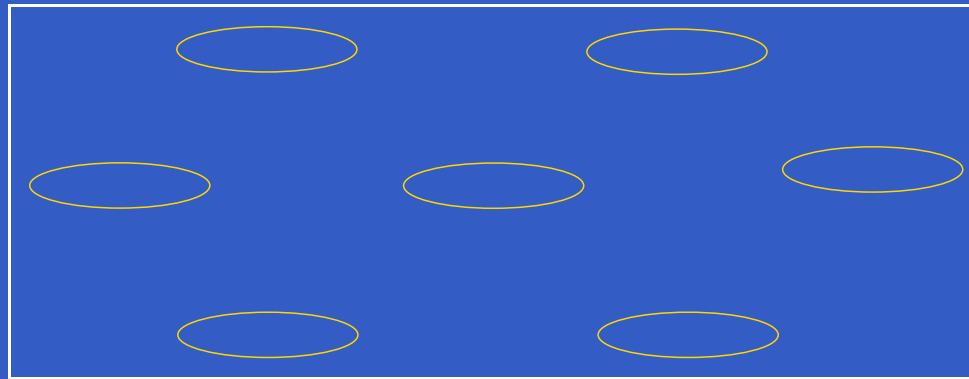
$D_l \leftarrow$ *Muestrear* M individuos de $p_l(\mathbf{x})$

- No hay cruce ni mutación. Generación descendientes \rightarrow simulación distribución de probabilidad.
- Similitudes con LEM ('Learnable Evolution Model', Michalski'00).

EDAs: división por grados de complejidad

- Punto de flotación de los EDAs → estimación de la distribución de probabilidad de los individuos seleccionados.
- Literatura → aproximaciones con distintos *grados de complejidad* del modelo probabilístico:
 - Variables del problema independientes.
 - Captura de relaciones bivariadas.
 - Captura de relaciones multivariadas.
- Simulación del modelo probabilístico: PLS ('Probabilistic Logic Sampling', Henrion'88).

EDAs: dominios discretos, variables independientes



$$p_l(\mathbf{x}) = p(\mathbf{x} | D_{l-1}^{Se}) = \prod_{i=1}^n p_l(x_i).$$

- UMDA ('Univariate Marginal Distribution Algorithm', Mühlenbein'98):

$$p_l(x_i) = \frac{\sum_{j=1}^N \delta_j(X_i=x_i | D_{l-1}^{Se})}{N}.$$

- BSC ('Bit-Based Simulation Crossover', Syswerda'93):

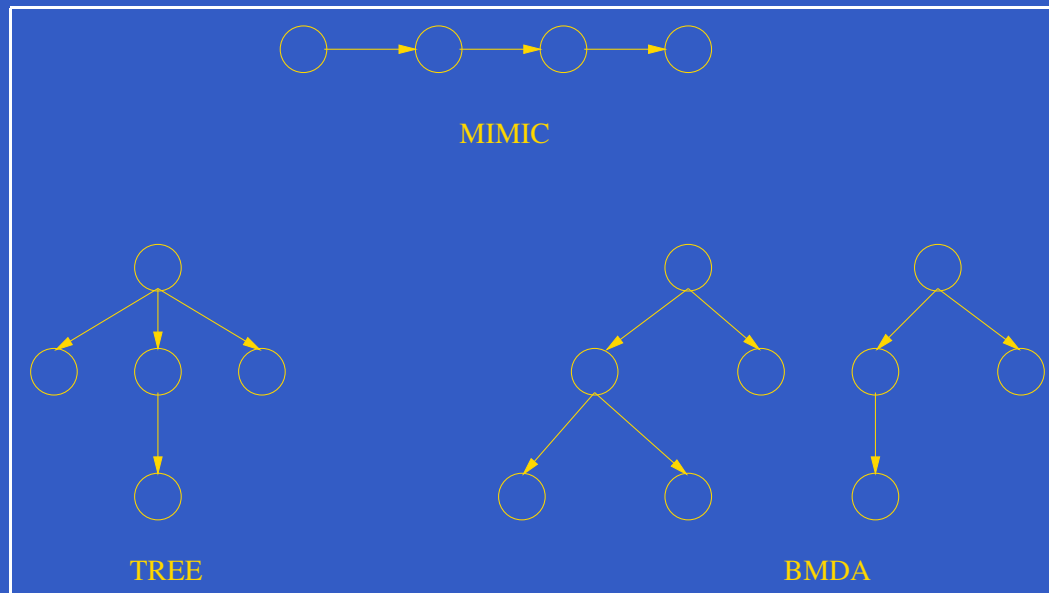
$$p_l(x_i) = \frac{\sum_{\{\mathbf{x} | \delta_j(X_i=x_i | D_{l-1}^{Se})=1\}} ef(\mathbf{x})}{\sum_{\{\mathbf{x} \in D_{l-1}^{Se}\}} ef(\mathbf{x})}.$$

- PBIL ('Population Based Incremental Learning', Baluja'94):

$$p_l(\mathbf{x}) = (1 - \alpha)p_{l-1}(\mathbf{x}) + \alpha \frac{1}{N} \sum_{k=1}^N \mathbf{x}_{k:M}^{l-1}.$$

- cGA ('compact Genetic Algorithm', Harik y col.'98).

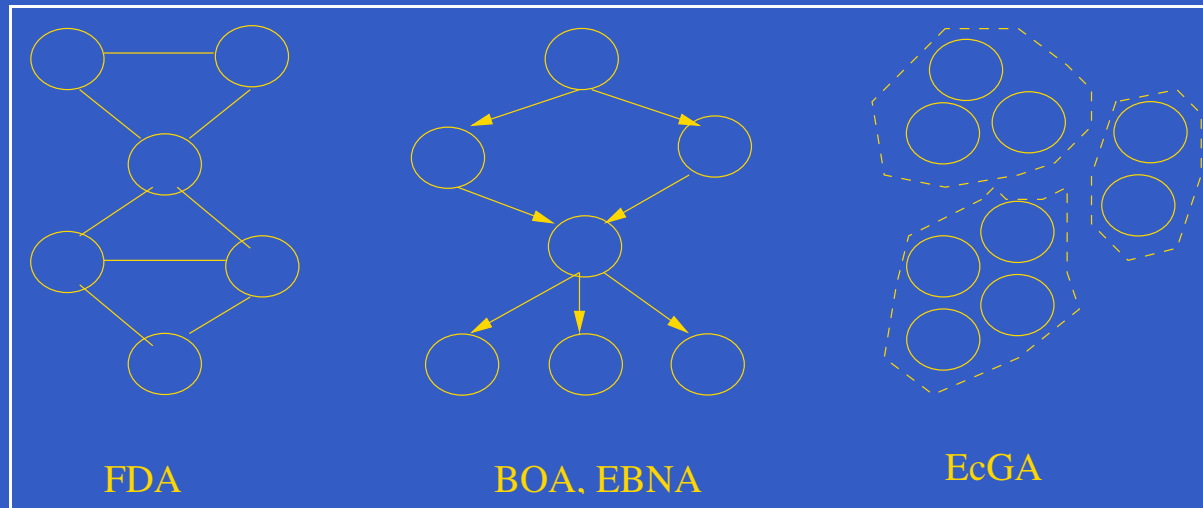
EDAs: dominios discretos, dependencias bivariadas



- MIMIC ('Mutual Information Maximization for Input Clustering', De Bonet y col.'97):

$$p_l^\pi(\mathbf{x}) = p_l(x_{i_1} | x_{i_2}) \cdot p_l(x_{i_2} | x_{i_3}) \cdot \dots \cdot p_l(x_{i_{n-1}} | x_{i_n}) \cdot p_l(x_{i_n})$$
- TREE (Chow y Liu'68): única estructura de árbol
$$p_l^t(\mathbf{x}) = \prod_{i=1}^n p_l(x_i | x_{j(i)})$$
.
- BMDA ('Bivariate Marginal Distribution Algorithm', Pelikan y Mühlenbein'99): conjunto de estructuras de árbol.

EDAs: dominios discretos, relaciones multivariadas



- FDA ('Factorized Distribution Algorithm', Mühlenbein y col.'99): factorización fija de la distribución, dada por el experto.
- BOA ('Bayesian Optimization Algorithm', Pelikan'99): red Bayesiana ('score' → BDe, 'search' → Algoritmo B).
- EBNA ('Estimation of Bayesian Networks Algorithm', Etxeberria y Larrañaga'99): red Bayesiana ('score' → BIC, 'search' → Algoritmo B).
- EcGA ('Extended compact Genetic Algorithm', Harik'99).

EDAs: dominios continuos

- Variables independientes: UMDA_c (Larrañaga y col.'99), SHCLVND (Rudlof y Köppen'96), PBIL_c (Sebag y Ducoulombier'98).
- Dependencias bivariadas: MIMIC_c^G (Larrañaga y col.'99), funciones Gaussianas bivariadas.
- Dependencias múltiples:
 - Funciones multivariadas de densidad normal (Larrañaga y col.'99).
 - EGNA ('Estimation of Gaussian Networks Algorithm', Larrañaga y col.'99): redes Gaussianas ('score' → BGe, 'search' → Algoritmo B).

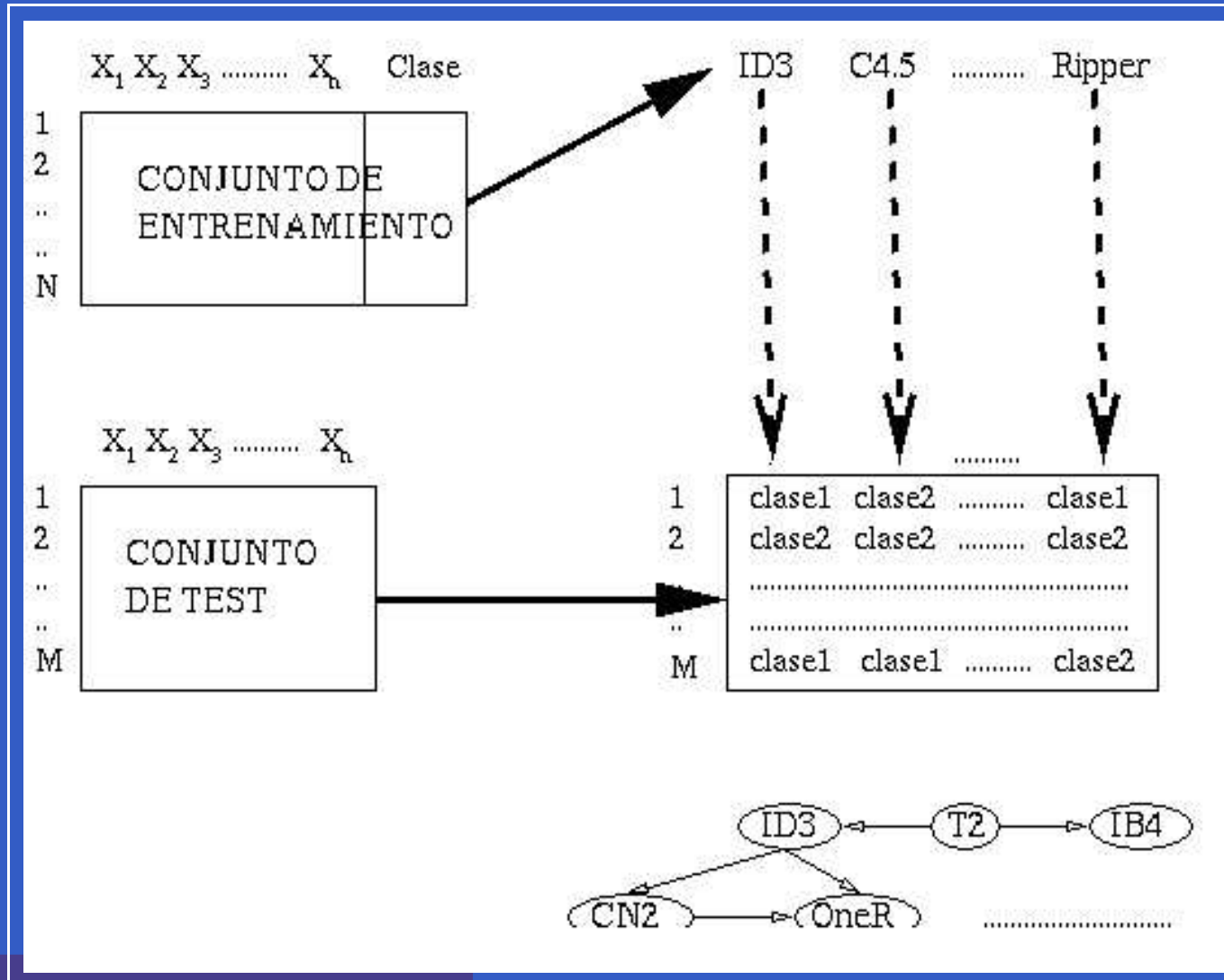
Comportamiento conjunto clasificadores supervisados

- ¿Algoritmo de clasificación adecuado para los datos?
- Comunidad de investigadores en Clasificación Supervisada → *porcentaje de bien clasificados*.
- Nuestro objetivo:
 - Estudiar el *comportamiento conjunto* de los clasificadores.
 - Estudiar las *relaciones entre* los clasificadores.
- Nos basaremos → *etiquetas de clase predecidas* por los clasificadores → naturaleza del modelo clasificadorio.
- Etiquetas de clase predecidas → redes Bayesianas → estudio comportamiento conjunto clasificadores y relaciones.

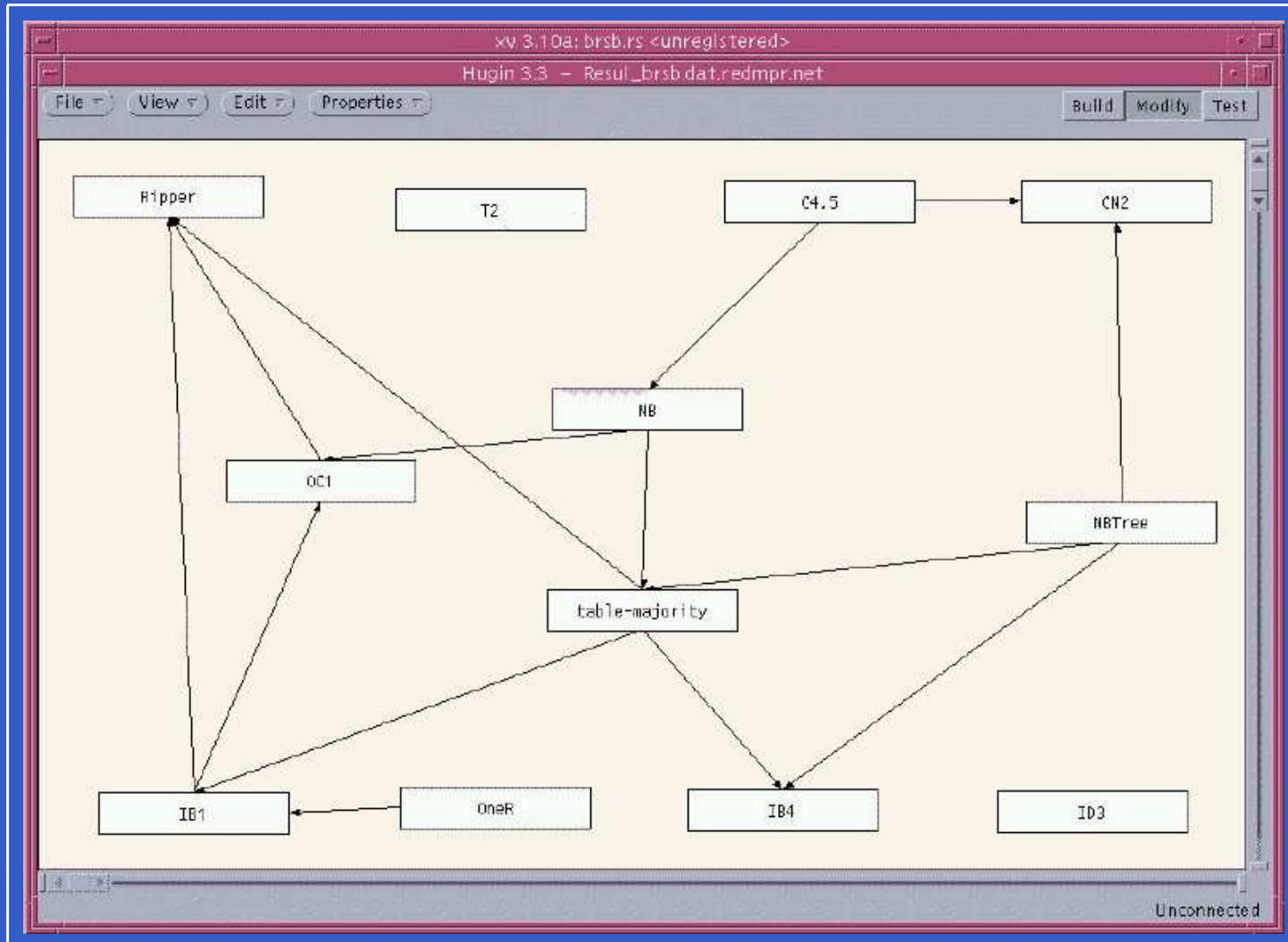
Comportamiento conjunto: clasificadores y dominios

- 14 algoritmos de clasificación supervisada:
 - Familia de Arboles de clasificación: C4.5, ID3.
 - Familia de K-NN: IB1, IB4.
 - Familia de Reglas de decisión IF-THEN: CN2, Ripper.
 - Familia de Arboles simples: OneR, T2.
 - Familia de clasificadores Bayesianos simples: Naive-Bayes (NB), NBTree.
 - Otros: HOODG, OC1, PEBLS, Table-Majority.
- 11 bases de datos médicas del repositorio UCI: *Breast cancer, Breast cancer (Wisconsin), Cleveland, Diabetes (Pima), Echocardiogram, Heart disease, Hepatitis, Hungarian, Hypothyroid, Liver (BUPA) y Lymphography.*

Comportamiento conjunto: metodología



Ejemplo: Breast Cancer



Comportamiento conjunto clasificadores supervisados

- Finalmente → 11 redes Bayesianas, una por dominio.
- 3 variantes del concepto de independencia condicional → indagar en el comportamiento conjunto de los clasificadores y relaciones entre ellos.
- Basándonos en el número de dominios (redes Bayesianas) que un clasificador (o familia de clasificadores) muestra las variaciones de independencia condicional propuestas → extraeremos tendencias o pautas de comportamiento de los clasificadores (o familias).

Independencia condicional fuerte

- Un clasificador (X) presenta la propiedad de *independencia condicional fuerte* si, dado otro clasificador (Y), es condicionalmente independiente al resto de clasificadores: $I(X, \mathbf{W} \setminus \{X, Z\} | Y)$ para cualquier clasificador Z .
- Modelo clasificatorio de $X \rightarrow$ 'original' o 'diferente'.
- Tendencia de: IB1, IB4, ID3, OneR, Table-Majority, T2.

Independencia condicional dentro de una familia

- Dos clasificadores de la misma familia (X, Y) , ¿son condicionalmente independientes dado otro clasificador (Z) ?
¿ $I(X, Y | Z)$ o $D(X, Y | Z)$?
- Grado de disimilaridad (similaridad) entre los clasificadores de una misma familia.
- Tendencia:
 - Hacia la disimilaridad: Árboles de clasificación, Reglas de decisión IF-THEN, Árboles simples.
 - Hacia la similaridad: K-NN, clasificadores Bayesianos simples.

Independencia condicional entre familias

- Estudio de la independencia condicional entre los clasificadores de distintas familias (\mathbf{X} , \mathbf{Y}), dado cualquier otro grupo de clasificadores (\mathbf{Z}): ¿ $I(\mathbf{X}, \mathbf{Y} | \mathbf{Z})$ o $D(\mathbf{X}, \mathbf{Y} | \mathbf{Z})$?
- Grado de disimilaridad (similaridad) entre distintas familias de clasificadores.
- Tendencias remarcables:
 - Hacia la disimilaridad:
clasificadores Bayesianos simples vs. Arboles simples.
Arboles de clasificación vs. K-NN.
 - Hacia la similaridad:
Arboles de clasificación vs. Reglas de decisión IF-THEN.

Trabajo futuro

- Distintas metodologías: clustering jerárquico, tests estadísticos.
- Relación de este tipo de propiedades con la usual mejora clasificatoria de la *combinación de clasificadores*.
- Trabajos similares → redes Genéticas → relaciones entre genes.

Selección de variables: motivación

$$X_1, X_2, \dots, X_n$$

- ¿Son todas las variables *útiles* para el aprendizaje del modelo clasificadorio?
- No monotonidad: porcentaje de bien clasificados vs. adición de variables.
 - Irrelevantes: $X_i \perp Clase.$
 - Redundantes: $X_i \equiv X_j.$
- Selección de variables: ‘dado un grupo de variables, escoger el mejor subconjunto de ellas para un problema de clasificación’.
- FSS \equiv ‘Feature Subset Selection’.
- Ventajas que acarrea:
 - Reducción en el coste de adquisición de datos.
 - Facilitación de la comprensión del modelo clasificadorio.
 - Rapidez en la inducción del clasificador final.
 - *Mejora en la bondad clasificatoria.*

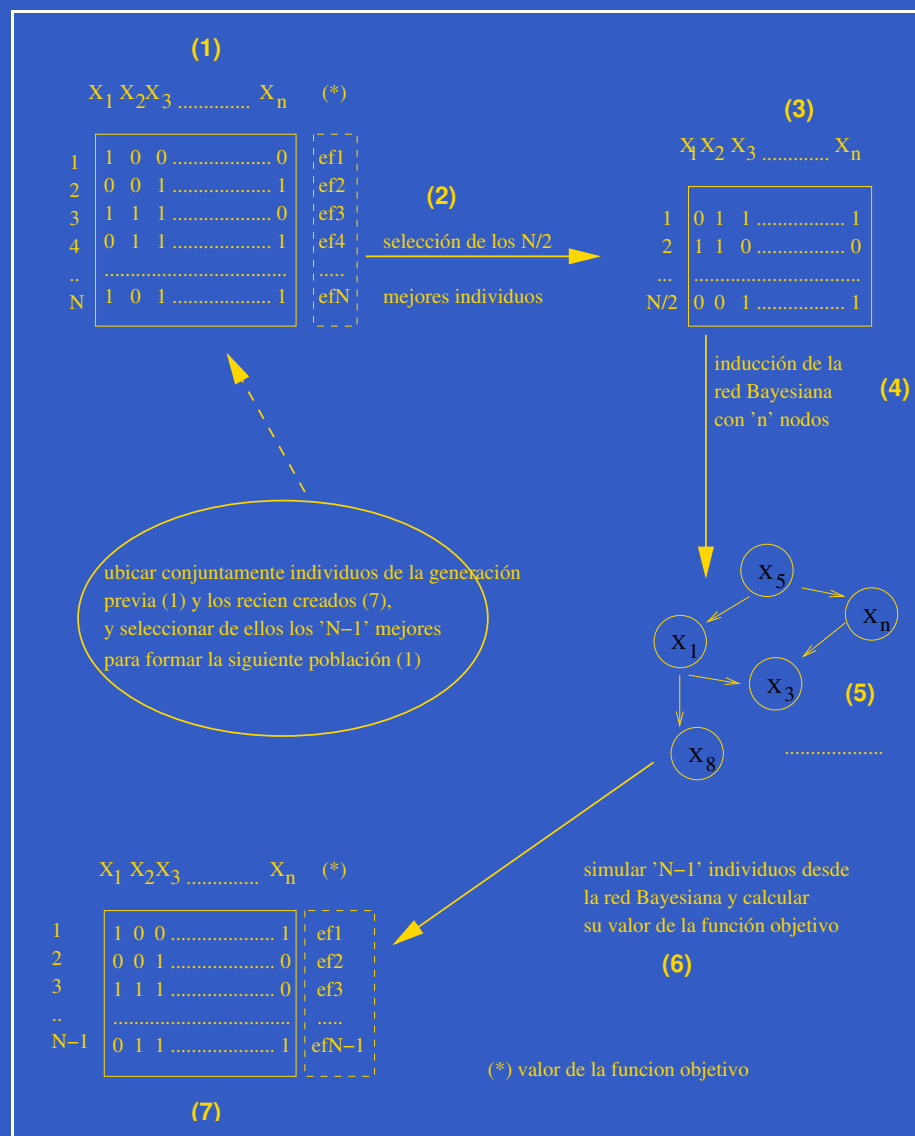
Selección de variables: problema de búsqueda

- Pattern Recognition, Bioinformática, Text-Learning, Clasificación no supervisada...
- Clásico problema NP-duro → heurísticos. Componentes del proceso de búsqueda:
 - Componente 1: punto de inicio de la búsqueda.
 - Componente 2: estrategia de la búsqueda.
 - Completa: anchura, profundidad, Branch and Bound.
 - Heurística: Simulated Annealing, Floating, Secuenciales, Algoritmos Genéticos...
 - Componente 3: cálculo de la función de evaluación.
 - 'Filter': características internas de los datos.
 - 'Wrapper' (John y col.'94): estimación de la bondad predictiva del clasificador a utilizar, usando únicamente las variables seleccionadas.
 - Componente 4: criterio de parada de la búsqueda.

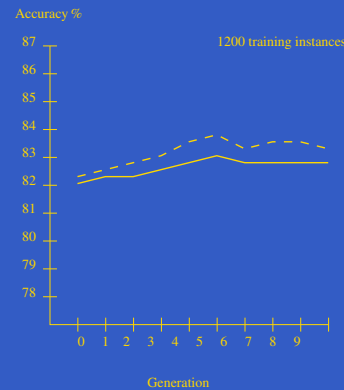
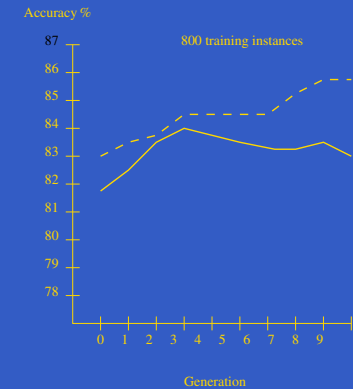
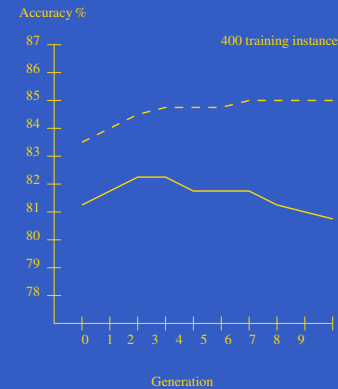
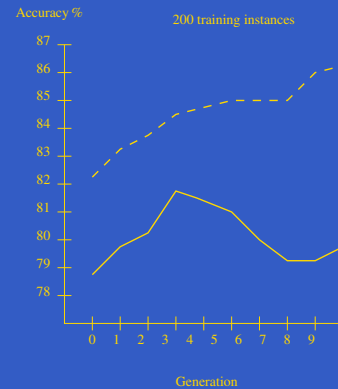
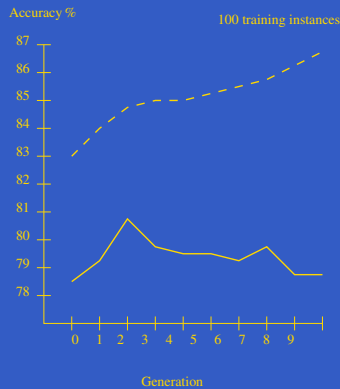
Selección de variables mediante EDAs

- Problemas de dimensionalidad baja (0-19) y media (20-49): EBNA.
- Problemas de dimensionalidad alta (>50): PBIL, BSC, MIMIC, TREE.
- Función de evaluación: 'wrapper' sobre Naive-Bayes, 10-fold cross-validation.

Selección de variables mediante EDAs



Selección de variables: sobrentrenamiento



- 'Overfitting' en selección de variables: Kohavi'95, Ng'97, ...
- Estudio FSS-EBNA → *Waveform-40*.
- Diferencia entre estimación sobre datos de entrenamiento y testeo.
- Mejoras sobre datos de entrenamiento → empeoramiento sobre datos de testeo.

Comparativa: baja y media dimensionalidad

<i>Dominio</i>	<i>Número de instancias</i>	<i>Número de variables</i>
<i>Tonosphere</i>	351	34
<i>Horse-colic</i>	368	22
<i>Soybean-large</i>	683	35
<i>Anneal</i>	898	38
<i>Image</i>	2,310	19
<i>Sick-euthyroid</i>	3,163	25

- SFS ('Sequential Forward Selection').
- SBE ('Sequential Backward Elimination').
- GA-o (Cruce en un punto).
- GA-u (Cruce uniforme).
- EBNA.

Comparativa: baja y media dimensionalidad

<i>Dominio</i>	<i>no FSS</i>	<i>SFS</i>	<i>SBE</i>
<i>Ionosphere</i>	84,84 ± 3,12†	90,25 ± 1,58*	91,39 ± 2,68
<i>Horse-colic</i>	78,97 ± 2,98†	83,31 ± 1,98	82,12 ± 2,41*
<i>Soybean-large</i>	81,96 ± 3,46†	86,38 ± 3,30*	87,78 ± 3,90*
<i>Anneal</i>	93,01 ± 3,13*	86,72 ± 2,09†	92,49 ± 2,94*
<i>Image</i>	79,95 ± 1,52†	88,65 ± 1,21	88,82 ± 1,74
<i>Sick-euthyroid</i>	84,77 ± 2,70†	90,73 ± 0,55†	95,57 ± 0,16
Media	83,91	87,67	89,69

<i>Dominio</i>	<i>FSS-GA-o</i>	<i>FSS-GA-u</i>	<i>FSS-EBNA</i>
<i>Ionosphere</i>	91,17 ± 3,19	90,97 ± 2,56*	92,40 ± 2,04
<i>Horse-colic</i>	83,43 ± 2,82	83,51 ± 1,47	83,93 ± 1,58
<i>Soybean-large</i>	85,64 ± 4,06†	86,09 ± 4,37†	88,64 ± 1,70
<i>Anneal</i>	92,95 ± 2,67*	93,13 ± 2,56	94,10 ± 3,00
<i>Image</i>	88,67 ± 2,48	89,12 ± 1,56	88,98 ± 0,98
<i>Sick-euthyroid</i>	95,97 ± 0,58	95,90 ± 0,43	96,14 ± 0,65
Media	89,63	89,78	90,69

- Validación 5x2cv; †, (0,05); *, (0,1).
- Todos los algoritmos → Mejoras sobre la no selección de variables.
- Mejor comportamiento predictivo medio de FSS-EBNA.

Generación de parada: EDAs vs. genéticos

- Comparación EBNA vs. genéticos: similar bondad → ¿Generación de parada?

<i>Dominio</i>	<i>FSS-GA-o</i>	<i>FSS-GA-u</i>	<i>FSS-EBNA</i>
<i>Ionosphere</i>	3,50 ± 0,84†	3,10 ± 0,56†	1,80 ± 0,42
<i>Horse-colic</i>	3,20 ± 1,13*	3,40 ± 0,51†	2,40 ± 0,69
<i>Soybean-large</i>	3,30 ± 0,82*	3,60 ± 0,51†	2,50 ± 0,70
<i>Anneal</i>	3,80 ± 0,42†	3,20 ± 0,44†	1,80 ± 0,42
<i>Image</i>	3,60 ± 0,84	3,70 ± 0,48	3,50 ± 0,42
<i>Sick-euthyroid</i>	4,50 ± 0,70*	4,80 ± 0,42*	3,50 ± 0,97

- Similar bondad clasificatoria → EBNA necesita menos generaciones para alcanzar similares niveles de bondad.
- Parada ‘prematura’ en similares niveles de bondad → muy deseable con funciones de evaluación ‘wrapper’ (1” *Ionosphere*, 3” *Image*) → ahorro del cálculo de nuevas generaciones.

Generación de parada: EDAs vs. genéticos

<i>Dominio</i>	<i>FSS-GA-o</i>	<i>FSS-GA-u</i>	<i>FSS-EBNA</i>
<i>Redundant-order-3</i>	2,50 ± 1,76	3,33 ± 1,63	1,33 ± 0,81
<i>Redundant-order-5</i>	3,83 ± 2,04	3,33 ± 2,16	1,83 ± 0,75
<i>Redundant-order-7</i>	2,00 ± 1,67	2,50 ± 1,76	1,00 ± 1,09

- Confirmación sobre dominios artificiales.
- Deterioro del comportamiento de GA-o al no codificar juntas las variables relacionadas.

Comparativa: alta dimensionalidad

<i>Dominio</i>	<i>Número de instancias</i>	<i>Número de variables</i>
<i>Audiology</i>	226	69
<i>Arrhythmia</i>	452	279
<i>Cloud</i>	1,834	204
<i>DNA</i>	3,186	180
<i>Internet advertisements</i>	3,279	1,558
<i>Spambase</i>	4,601	57

- BSC, PBIL.
- MIMIC, TREE.
- GA-o (Cruce en un punto), GA-u (Cruce uniforme).

Comparativa: alta dimensionalidad

<i>Dominio</i>	<i>no FSS</i>	<i>FSS-GA-o</i>	<i>FSS-GA-u</i>
<i>Audiology</i>	52,39 ± 5,56†	68,29 ± 2,98	68,44 ± 4,46
<i>Arrhythmia</i>	39,91 ± 8,50†	63,23 ± 3,95	64,73 ± 3,52
<i>Cloud</i>	68,18 ± 2,09†	74,49 ± 1,93	75,17 ± 1,22
<i>DNA</i>	93,93 ± 0,67	94,00 ± 0,75	95,01 ± 0,56
<i>Internet advertisements</i>	95,23 ± 0,40*	96,10 ± 0,12	96,38 ± 0,47
<i>Spambase</i>	81,71 ± 0,92†	88,92 ± 1,45	88,77 ± 1,28
<i>Media</i>	71,88	80,83	81,41

<i>Dominio</i>	<i>FSS-PBIL</i>	<i>FSS-BSC</i>	<i>FSS-MIMIC</i>	<i>FSS-TREE</i>
<i>Audiology</i>	70,22 ± 2,78	68,29 ± 3,18	68,88 ± 3,93	70,09 ± 4,12
<i>Arrhythmia</i>	64,62 ± 2,70	65,01 ± 2,22	64,33 ± 1,82	64,51 ± 2,59
<i>Cloud</i>	75,18 ± 1,30	76,24 ± 1,25	76,31 ± 0,95	75,84 ± 0,98
<i>DNA</i>	94,86 ± 0,64	95,40 ± 0,40	95,53 ± 0,29	95,40 ± 0,28
<i>Internet adv.</i>	96,49 ± 0,21	96,37 ± 0,41	96,46 ± 0,46	96,69 ± 0,63
<i>Spambase</i>	88,63 ± 1,36	89,52 ± 1,38	89,80 ± 0,79	89,60 ± 0,93
<i>Media</i>	81,66	81,80	81,88	82,02

- Todos los algoritmos → Mejoras sobre la no selección.
- Sin diferencias significativas entre las bondades predictivas.

Generación de parada: EDAs vs. genéticos

- Similares bondades predictivas → ¿Generación de parada?.

<i>Dominio</i>	<i>FSS-GA-o</i>	<i>FSS-GA-u</i>	<i>FSS-PBIL</i>
<i>Audiol.</i>	5,80 ± 0,42†	4,60 ± 0,96*	5,20 ± 1,03*
<i>Arrhyt.</i>	8,70 ± 0,48†	8,80 ± 0,42†	8,30 ± 0,48*
<i>Cloud</i>	10,50 ± 0,52*	10,60 ± 1,07*	10,40 ± 0,84
<i>DNA</i>	12,80 ± 0,91†	11,80 ± 0,42†	11,30 ± 0,48†
<i>Int. adv.</i>	4,70 ± 1,41	5,00 ± 1,41	5,00 ± 0,66
<i>Spamb.</i>	4,80 ± 1,03	5,20 ± 0,63	5,50 ± 1,17

<i>Dominio</i>	<i>FSS-BSC</i>	<i>FSS-MIMIC</i>	<i>FSS-TREE</i>
<i>Audiol.</i>	2,50 ± 0,70	2,80 ± 0,78	2,80 ± 0,78
<i>Arrhyt.</i>	7,10 ± 0,73	7,00 ± 0,66	7,20 ± 0,78
<i>Cloud</i>	8,40 ± 0,51	8,40 ± 0,69	8,30 ± 0,82
<i>DNA</i>	8,70 ± 0,82	8,10 ± 0,73	8,40 ± 0,69
<i>Int. adv.</i>	4,40 ± 1,26	4,30 ± 0,67	4,00 ± 1,63
<i>Spamb.</i>	4,20 ± 0,91	3,70 ± 0,82	4,20 ± 1,22

Generación de parada: EDAs vs. genéticos

- Similares bondades predictivas → dos tipos de comportamientos:
 - BSC, MIMIC, TREE: menor número de generaciones.
 - PBIL, GA-o, GA-u.
- Parada 'prematura' en similares niveles de bondad → muy deseable con funciones de evaluación 'wrapper' (1" *Audiology*, 9.8" *Internet advertisement*) → 'ahorro' del cálculo de nuevas generaciones.

Generación de parada: EDAs vs. genéticos

<i>Dominio</i>	<i>FSS-GA-o</i>	<i>FSS-GA-u</i>	<i>FSS-PBIL</i>
<i>Red60of1</i>	$6,70 \pm 0,48^\dagger$	$4,10 \pm 0,31$	$12,80 \pm 0,91^\dagger$
<i>Red30of2</i>	$22,40 \pm 4,22^\dagger$	$73,50 \pm 5,73^\dagger$	$66,30 \pm 7,52^\dagger$
<i>Red30of3</i>	$21,00 \pm 2,26^\dagger$	$119,00 \pm 5,27^\dagger$	$113,80 \pm 8,76^\dagger$

<i>Dominio</i>	<i>FSS-BSC</i>	<i>FSS-MIMIC</i>	<i>FSS-TREE</i>
<i>Red60of1</i>	$7,60 \pm 0,51^\dagger$	$8,60 \pm 0,51^\dagger$	$8,00 \pm 0,47^\dagger$
<i>Red30of2</i>	$36,40 \pm 3,13^\dagger$	$15,10 \pm 2,33$	$10,90 \pm 1,52$
<i>Red30of3</i>	$89,50 \pm 17,60^\dagger$	$18,90 \pm 2,13$	$16,30 \pm 1,33$

- Confirmación sobre dominios artificiales.
- Deterioro del comportamiento de GA-o al no codificar juntas las variables relacionadas.

Trabajo futuro

- Aprendizaje paralelo de redes Bayesianas en dominios de alta dimensionalidad, dentro del paradigma EDA.
- Selección de genes en dominios de *DNA microarrays* → valor numérico de la expresión conjunta de miles de genes.

Pesado de atributos para K-NN: 1-NN

- Clasificador 1-NN (vecino más próximo), para clasificar una instancia de test:
 - Guardar todas las instancias de entrenamiento.
 - Computar las *distancias* entre la instancia de test y cada instancia de entrenamiento: cálculo de *diferencias* en cada dimensión y suma de éstas:

$$distancia(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^n diferencia(x_i, y_i)^2}$$

- Clasificar la instancia de test con la clase de la instancia de entrenamiento más *cercana*.
- Las *diferencias* en cada dimensión son sumadas de una manera '*ingenua*' → otorgando la misma relevancia (o peso) a todos los atributos → pesos homogéneos en todas las dimensiones.

Pesado de atributos para K-NN: problema de búsqueda

- Propuesta ‘menos ingenua’ → mejora bondad clasificatoria:

$$distancia(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^n w_i \times diferencia(x_i, y_i)^2}$$

w_i : peso o relevancia del i -ésimo atributo.

- Problema NP-duro de búsqueda de los pesos óptimos (FW ≡ ‘Feature Weighting’): genéticos, secuenciales,...
- ¿Espacio de búsqueda? Pesos continuos vs. Pesos discretos → ¿Cuántos?.
- Kohavi y col.’97:
 - No monotonicidad: porcentaje de bien clasificados vs. número de pesos distintos.
 - No recomienda más de tres pesos discretos: $\{0, 0.5, 1\}$.

Pesado de atributos para K-NN mediante EDAs

- FW-EBNA: búsqueda en el espacio discreto de tres pesos: $\{0, 0.5, 1\}$.
- FW-EGNA: búsqueda en el espacio continuo de pesos: $[0, 1]$.
- Función de evaluación: 'wrapper', estimación mediante 'leave-one-out'.

Pesado de atributos para K-NN: comparativa

- FW-EBNA (3 pesos discretos)
- FW-EGNA, ([0, 1])
- FW-GA-o (3 pesos discretos)
- DIET ('Best-First', Kohavi y col.'97, 3 pesos discretos)
- IB4 ('hill-climbing', Aha'92, [0, 1])
 - Excepto IB4 ('filter'), todos los algoritmos función de evaluación 'wrapper' ('leave-one-out').

<i>Dominio</i>	<i>Número de instancias</i>	<i>Número de atributos</i>
(1) <i>LED24</i>	600	24
(2) <i>Waveform-21</i>	600	21
(3) <i>3-Weights</i>	600	12
(4) <i>C-Weights</i>	600	10
(5) <i>Glass</i>	214	9
(6) <i>CRX</i>	690	15
(7) <i>Vehicle</i>	846	18
(8) <i>Contraceptive</i>	1,473	9

Pesado de atributos para K-NN: comparativa

<i>Dominio</i>	<i>no-FW</i>	<i>DIET-10</i>	<i>FW-GA-o</i>
<i>LED24</i>	47,37 ± 3,36†	63,84 ± 2,42*	68,64 ± 1,30
<i>Wave-21</i>	76,20 ± 1,48	76,66 ± 1,62	76,71 ± 1,57
<i>3-Weights</i>	77,19 ± 3,36†	81,91 ± 1,98*	82,88 ± 1,66
<i>C-Weights</i>	81,01 ± 1,14†	83,55 ± 1,56	83,93 ± 1,24
<i>Glass</i>	64,85 ± 2,15†	71,34 ± 4,89	71,32 ± 2,97
<i>CRX</i>	81,56 ± 1,92*	82,12 ± 2,01*	83,14 ± 1,81
<i>Vehicle</i>	67,33 ± 2,11	68,71 ± 1,48	69,86 ± 1,42
<i>Contraceptive</i>	43,61 ± 0,97†	47,54 ± 2,99	48,10 ± 2,50
Media artificiales	70,44	76,49	78,04
Media reales	64,33	67,42	68,10

<i>Dominio</i>	<i>FW-EBNA</i>	<i>IB4</i>	<i>FW-EGNA</i>
<i>LED24</i>	69,03 ± 1,54	66,70 ± 1,80	61,55 ± 1,90†
<i>Wave-21</i>	76,87 ± 1,04	77,96 ± 1,62	76,90 ± 1,48
<i>3-Weights</i>	85,99 ± 1,57	80,32 ± 4,46†	82,00 ± 2,72
<i>C-Weights</i>	83,55 ± 1,56	81,98 ± 1,84†	84,33 ± 1,31
<i>Glass</i>	71,12 ± 5,01	61,13 ± 5,55*	70,09 ± 2,83
<i>CRX</i>	83,74 ± 1,94	85,48 ± 0,92	82,17 ± 2,14*
<i>Vehicle</i>	69,43 ± 2,11	64,65 ± 2,32*	69,58 ± 2,33
<i>Contraceptive</i>	48,32 ± 2,34	45,66 ± 2,66*	44,95 ± 2,10†
Media artificiales	78,86	76,74	76,19
Media reales	68,15	64,23	66,69

Pesado de atributos para K-NN: comparativa

- Todos los algoritmos → Mejoras sobre el no-pesado de atributos.
- Mejor comportamiento predictivo medio de FW-EBNA.
- Buen comportamiento general de DIET.
- FW-EGNA:
 - Corroboración de las tesis de Kohavi y col.'97.
 - No por ofrecer una 'amplia' gama de pesos → No aporta mejor bondad predictiva.
- Comparación EBNA vs. genético: similar bondad en la mayoría de dominios → ¿Generación de parada?

Generación de parada: EDAs vs. genéticos

<i>Domain</i>	<i>FW-GA-o</i>	<i>FW-EBNA</i>
<i>LED24</i>	6,50 ± 0,54†	4,50 ± 0,47
<i>Waveform-21</i>	2,50 ± 1,37	2,33 ± 1,03
<i>3-Weights</i>	2,50 ± 1,22	2,66 ± 0,81
<i>C-Weights</i>	5,16 ± 1,16	3,33 ± 1,50
<i>Glass</i>	1,00 ± 0,89	1,00 ± 0,63
<i>CRX</i>	1,33 ± 1,75	1,33 ± 0,51
<i>Vehicle</i>	3,16 ± 0,75*	1,16 ± 0,40
<i>Contraceptive</i>	2,66 ± 0,81*	1,83 ± 0,98

- Similar bondad clasificatoria → EBNA necesita menos generaciones para alcanzar similares niveles de bondad.
- Parada ‘prematura’ en similares niveles de bondad → muy deseable con funciones de evaluación ‘wrapper’ (3.3” *Glass*, 29.8” *LED24*, 60.4” *Vehicle*) → ahorro del cálculo de nuevas generaciones.

Trabajo futuro

- Búsqueda *conjunta*: conjunto óptimo de pesos y selección de prototipos mediante EDAs.
- Aplicación del K-NN en *DNA microarrays* → Búsqueda pesos óptimos → Nivel de relevancia de los genes en 3 estados.

Recopilatorio de conclusiones

- Estudio del comportamiento conjunto de clasificadores y sus relaciones:
 - Mediante redes Bayesianas.
 - 3 variantes del concepto de independencia condicional.
- Selección de variables en clasificación supervisada:
 - Propuesta de una batería de algoritmos EDA dependiendo de la dimensionalidad del dominio.
 - Estudio del fenómeno del 'overfitting' en la selección de variables.
 - Comparativa EDAs vs. genéticos → Número de generación en el que alcanzan similares niveles de bondad clasificatoria.
- Pesado de atributos del algoritmo del vecino más próximo:
 - Propuesta de dos nuevos algoritmos EDAs (redes Bayesianas y Gaussianas).
 - Comparativa EDAs vs. genéticos → Número de generación en el que alcanzan similares niveles de bondad clasificatoria.

Publicaciones que acompañan este trabajo: revistas internacionales

- I. Inza, B.Sierra, R. Blanco, P. Larrañaga (2002). Gene selection by sequential wrapper approaches in microarray cancer class prediction. **Journal of Intelligent and Fuzzy Systems**.
- I. Inza, M. Merino, P. Larrañaga, J. Quiroga, B. Sierra, M. Giralá (2001). Feature subset selection by genetic algorithms and estimation of distribution algorithms. A case study in the survival of cirrhotic patients treated with TIPS. **Artificial Intelligence in Medicine**, 23/2, 187-205.
- I. Inza, P. Larrañaga, B. Sierra (2001). Feature Subset Selection by Bayesian networks: a comparison with genetic and sequential algorithms. **International Journal of Approximate Reasoning**, 27/2, 143-164.
- I. Inza, P. Larrañaga, R. Etxeberria, B. Sierra (2000). Feature Subset Selection by Bayesian networks based optimization. **Artificial Intelligence**, 123(1-2),157-184.
- I. Inza, P. Larrañaga, B. Sierra, R. Etxeberria, J.A. Lozano, J.M. Peña (1999). Representing the joint behaviour of machine learning inducers by Bayesian networks. **Pattern Recognition Letters**, 20 (11-13), 1201-1209.

Publicaciones que acompañan este trabajo: capítulos de libro

- I. Inza, P. Larrañaga, B. Sierra (2001). Estimation of Distribution Algorithms for feature subset selection in large dimensionality domains. **Data Mining: A Heuristic Approach**. H. Abbass, R. Sarker, C. Newton (eds.), IDEA Group Publishing, 97-116.
- I. Inza, P. Larrañaga, B. Sierra (2001). Feature Subset Selection by Estimation of Distribution Algorithms. **Estimation of Distribution Algorithms. A new tool for Evolutionary Computation**. P. Larrañaga, J.A. Lozano (eds.), Kluwer Academic Publishers, 269-294.
- I. Inza, P. Larrañaga, B. Sierra (2001). Feature Weighting for Nearest Neighbor by Estimation of Distribution Algorithms. **Estimation of Distribution Algorithms. A new tool for Evolutionary Computation**. P. Larrañaga, J.A. Lozano (eds.), Kluwer Academic Publishers, 295-312.
- I. Inza, M. Merino, P. Larrañaga, J. Quiroga, B. Sierra, M. Giralá (2000). Feature Subset Selection Using Probabilistic Tree Structures. **Lecture Notes in Computer Sciences 1933**. R. Brause, E. Hanisch (eds.), Springer-Verlag, 97-110.