

Inferencia Estocástica y Aplicaciones de los Lenguajes de Árboles



Universitat d'Alacant
Universidad de Alicante

Autor: *Juan Ramón Rico*

Tesis doctoral

Dirigida por:

Dr. Rafael C. Carrasco

Dr. Jorge Calera

Introducción

Aprendizaje inductivo

Tarea de descubrir estructuras comunes a partir de ejemplos

Inferencia gramatical

Basada en lenguajes formales (Fu 1982)

Inferencia gramatical estocástica

Aporta información estadística

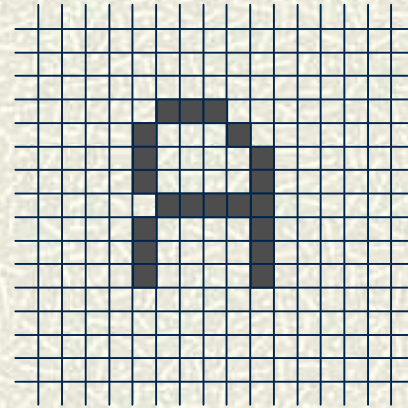
Introducción

Cadenas, árboles y grafos.

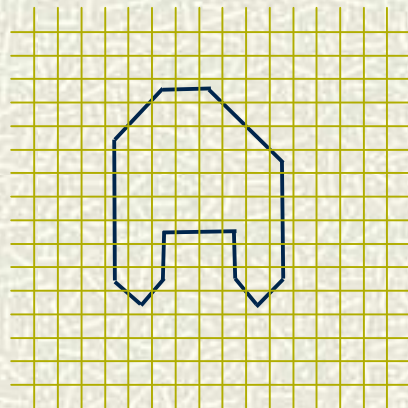
- Cadenas. Lenguajes de cadenas.
- Gramáticas de grafos
(Engelfriet y Heyker 1991; Fahmy y Blostein 1992; Courcelle et al. 1993).
 - Reconocimiento de patrones
(Flasinsky 1992 y Rekers 1994).
 - ✗ Problemas de eficiencia.
- Gramáticas de árboles
(Sima'an et al. 1996; Abe y Mamitsuka 1997; Carrasco et al. 2001).

Introducción

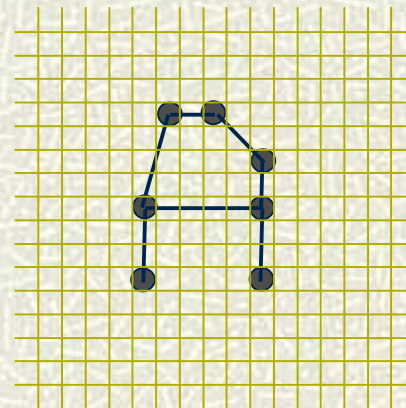
Entrada
datos



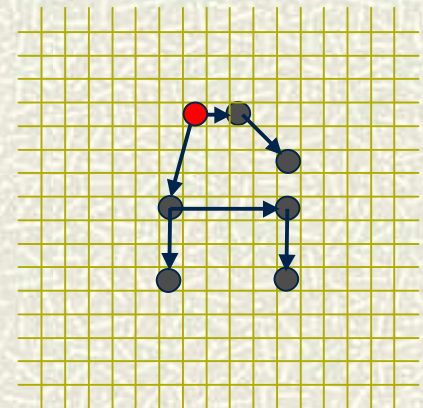
Representación
simbólica



cadena



grafo



árbol

Introducción

Compresión de texto.



	Estáticos	Adaptativos
Símbolo a símbolo	Huffman (5)	Compresión aritmética (2)
Basado en diccionario	-	Ziv-Lempel (4)

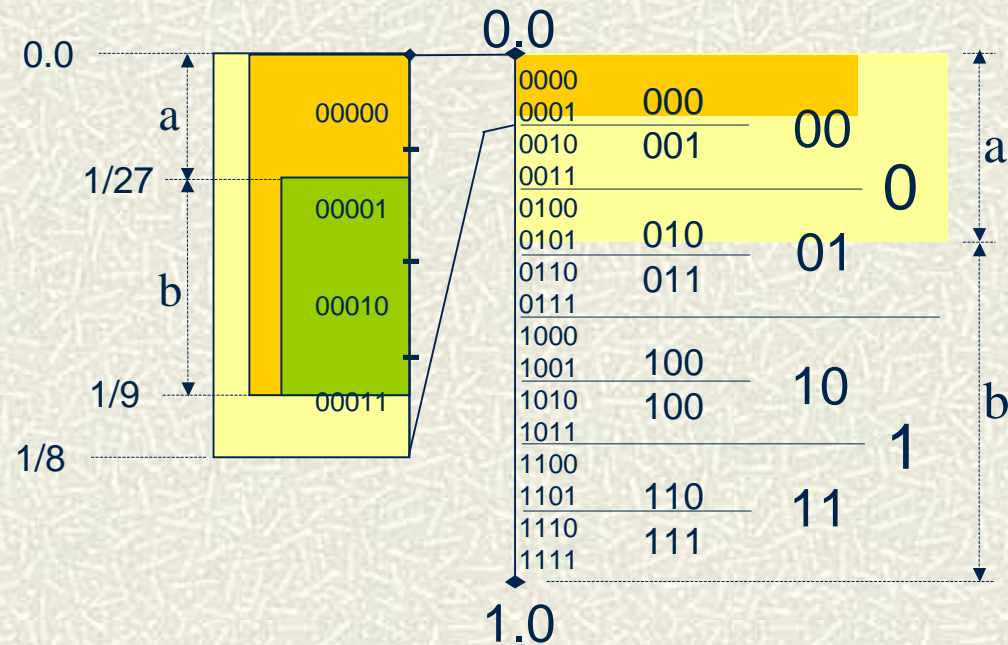
- ✓ Los mejores resultados se obtienen usando compresión aritmética

Codificación aritmética

$$\Sigma = \{a, b\}$$

$$p(a) = 1/3$$

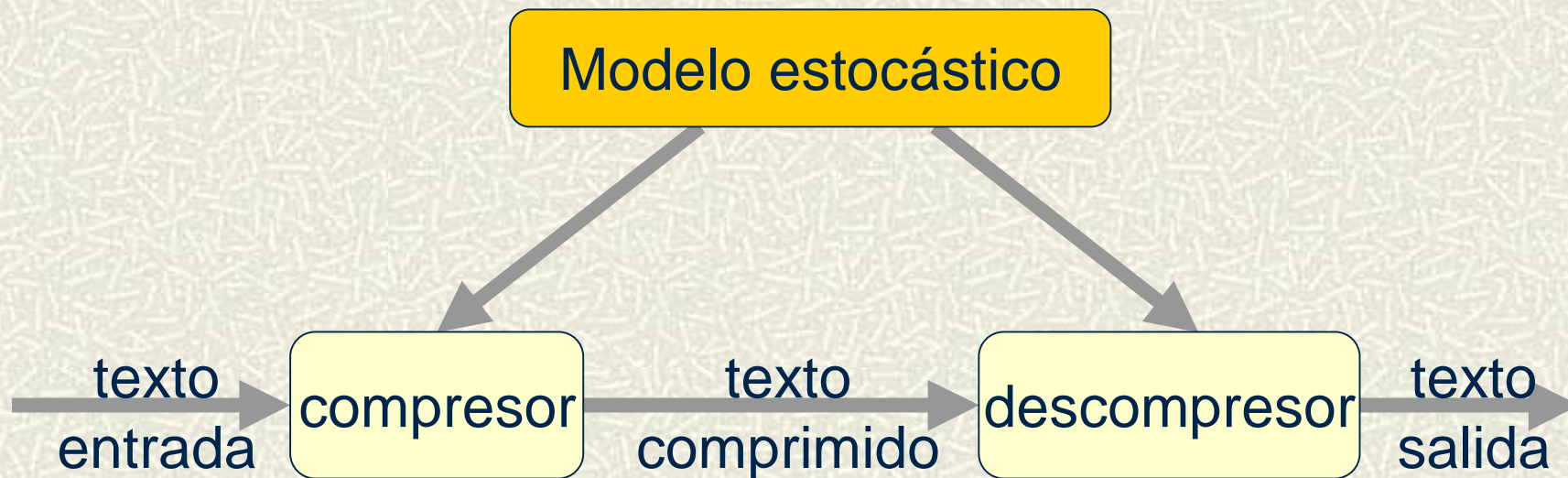
$$p(b) = 2/3$$



✗ No funciona con probabilidad nula

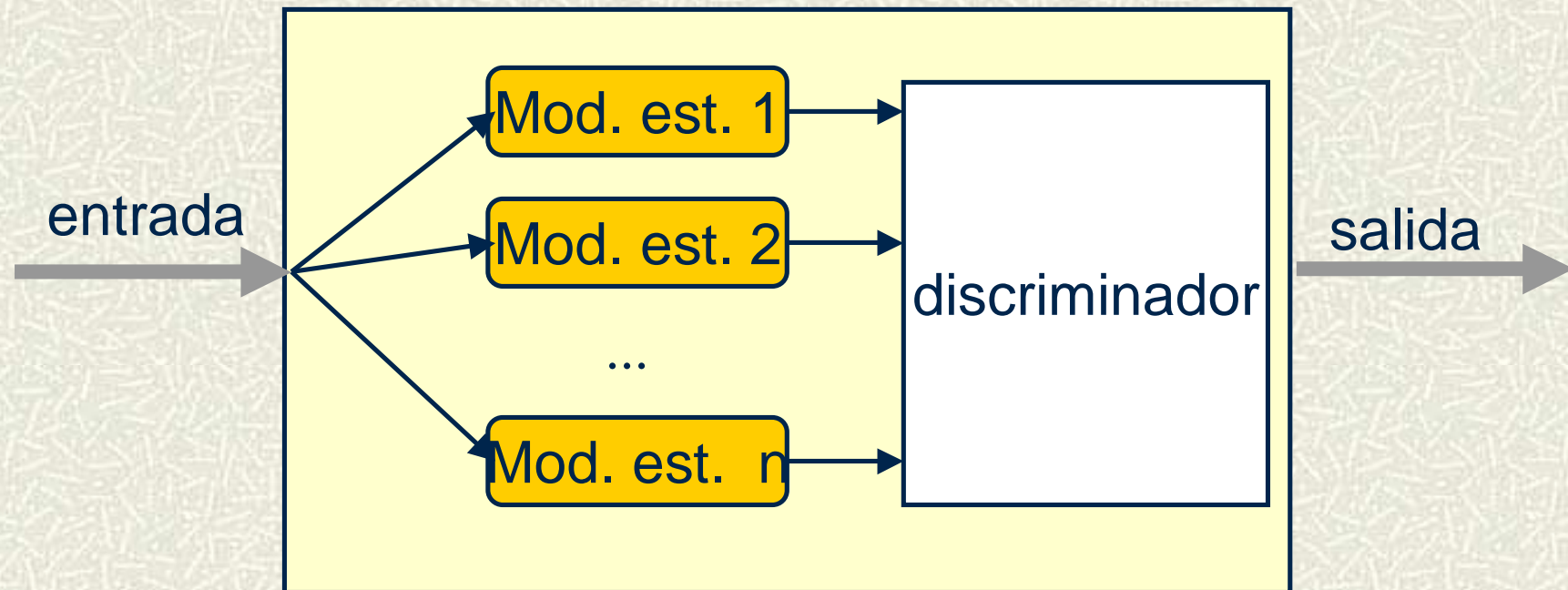
Modelos estocásticos para la compresión y la clasificación de cadenas

Esquema de compresión.



Modelos estocásticos para la compresión y la clasificación de cadenas

Esquema de clasificación.



Modelos estocásticos para la compresión y la clasificación de cadenas



Ventajas:

- ✓ Eficiencia.
- ✓ Incremental \rightarrow adaptativo.

Inconvenientes:

- ✗ Probabilidad nula (k grande) \rightarrow solución muy estudiada

Modelos estocásticos para la compresión y la clasificación de cadenas

Métodos de descuento básicos:

- Katz (1987).
- Absoluto (Ney y Essen 1993).
- Lineal (Katz 1987; Jelinek 1990).

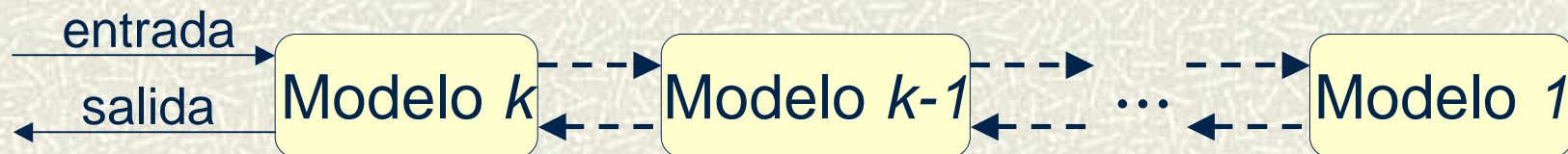
Métodos de descuento extendidos:

- Interpolación de modelos (Ney et al. 1997).
- Suavizado multinivel (Ney et al. 1997).

Modelos estocásticos para la compresión y la clasificación de cadenas

Modelos de predicción por concordancia parcial (PPM).

- Asignación de la probabilidad a eventos nos vistos.
- Compresión (Witten et al. 1999).



Parte I

Lenguajes de árboles k -testables

Modelos de contexto finito para árboles

- # Generalización natural de los k -gramas para lenguajes de árboles.
- # Estimación de las probabilidades a partir de las frecuencias experimentales.
- # Alternativa a métodos como:
 - *Inside-outside* (Sakakibara 1992).
 - Fusión de estados (Carrasco et al. 2001).

Modelos de contexto finito para árboles

	NO probabilísticos	probabilísticos
cadenas	<i>k</i> -testables (García y Vidal 1990; Yokomori 1995)	<i>k</i> -gramas
árboles	<i>k</i> -testables (Knuutila 1993; García 1993).	?

Modelos de contexto finito para árboles

Autómatas de árboles ascendentes deterministas (AAAD).

$$\Sigma = \{a, b\}$$

$$Q = \{q_1, q_2\}$$

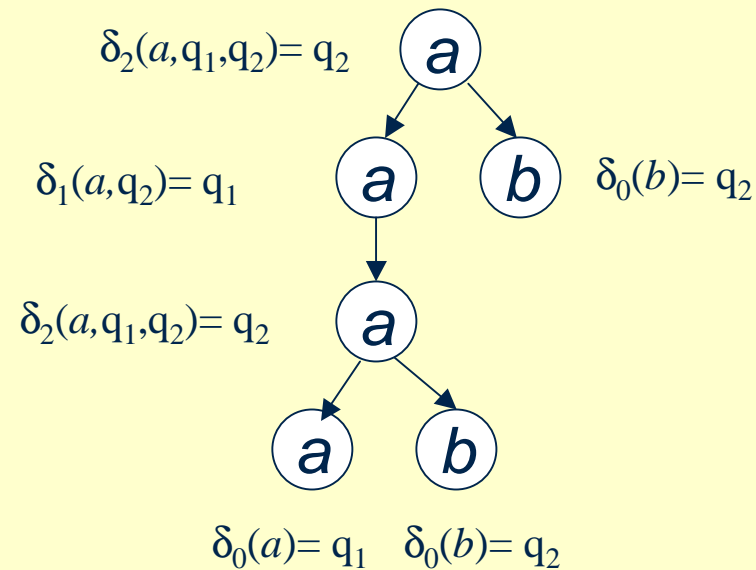
$$F = \{q_2\}$$

$$\delta_0(a) = q_1$$

$$\delta_0(b) = q_2$$

$$\delta_1(a, q_2) = q_1$$

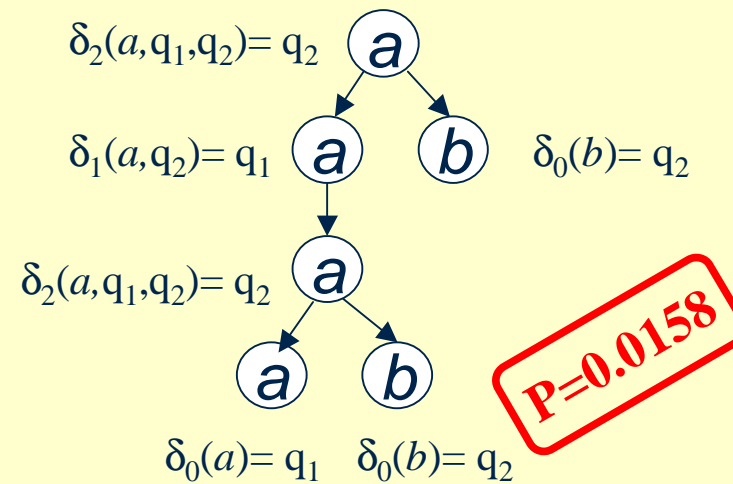
$$\delta_2(a, q_1, q_2) = q_2$$



Modelos de contexto finito para árboles

Autómatas de árboles ascendentes deterministas estocásticos (AAADE).

$\Sigma = \{a, b\}$	
$Q = \{q_1, q_2\}$	
$F = \{q_2\}$	1
$\delta_0(a) = q_1$	0.5
$\delta_0(b) = q_2$	0.3
$\delta_1(a, q_2) = q_1$	0.5
$\delta_2(a, q_1, q_2) = q_2$	0.7



Modelos de contexto finito para árboles

Autómatas k -testables.

- Definiciones: k -root, k -fork y k -subtree.

Para $k=3$ y $\Sigma=\{a,b\}$

$Q = \{ a, b, a(ab), a(a) \}$

$F = \{ a(ab) \}$

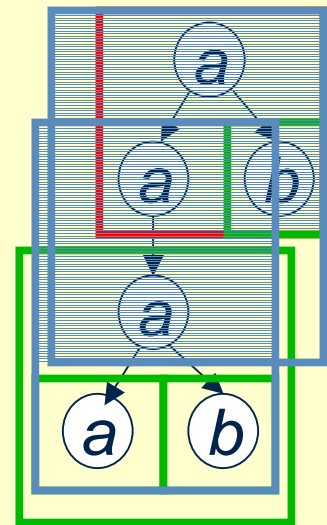
$\delta_0(a) = a$

$\delta_0(b) = b$

$\delta_1(a, a(ab)) = a(a)$

$\delta_2(a, a, b) = a(ab)$

$\delta_2(a, a(a), b) = a(ab)$

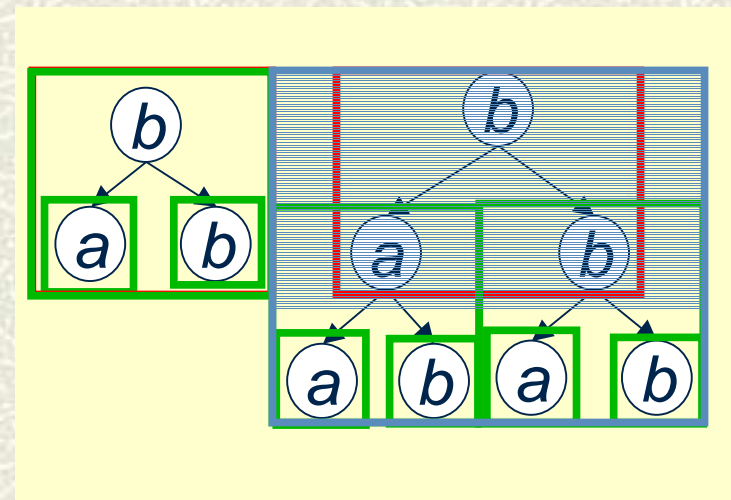


Modelos de contexto finito para árboles

Extensión estocástica de los lenguajes de árboles localmente testables

- Ejemplo: AAAD E para $k=3$

q	$r(q)$	$\sigma(t_1, \dots, t_m)$	$p_m(\sigma, t_1, \dots, t_m)$
a	0	a	$3/3$
b	0	b	$3/3$
$a(ab)$	0	$a(ab)$	$1/1$
$b(ab)$	$2/2$	$b(ab)$	$2/3$
		$b(a(ab)b(ab))$	$1/3$

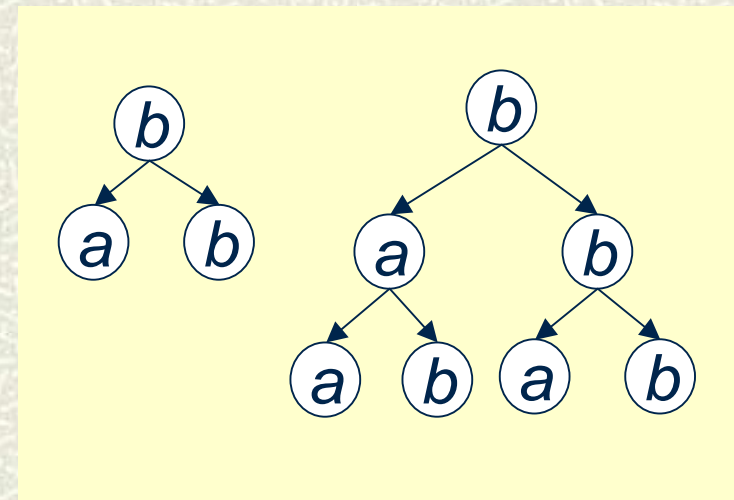


Modelos de contexto finito para árboles

Extensión estocástica de los lenguajes de árboles localmente testables

- Ejemplo modelos $M^{[3]}$ y $M^{[2]}$

$M^{[2]}$			
q	$r(q)$	$\sigma(t_1, \dots, t_m)$	$p_m(\sigma, t_1, \dots, t_m)$
a	0	a	$3/4$
		$a(ab)$	$1/4$
b	$2/2$	b	$3/6$
		$b(ab)$	$3/6$



Compresión mediante modelos k -testables adaptativos

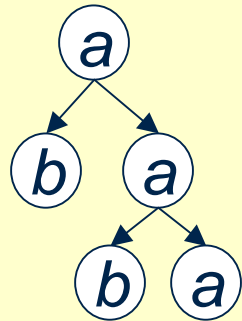
Método 1: modelo adaptativo con distribución a priori para árboles binarios.



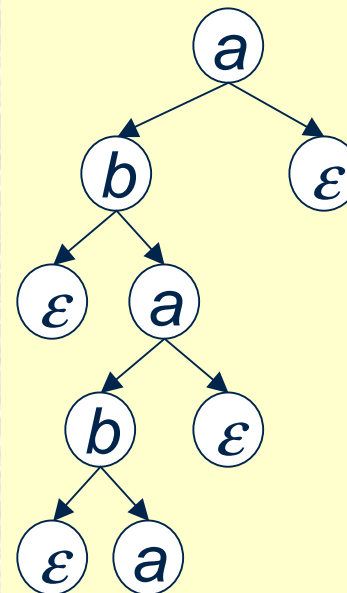
Compresión mediante modelos k -testables adaptativos

Método 1: modelo adaptativo con distribución a priori para árboles binarios.

n -ario



binario



Compresión mediante modelos k -testables adaptativos

Método 1: modelo adaptativo con distribución a priori para árboles binarios.

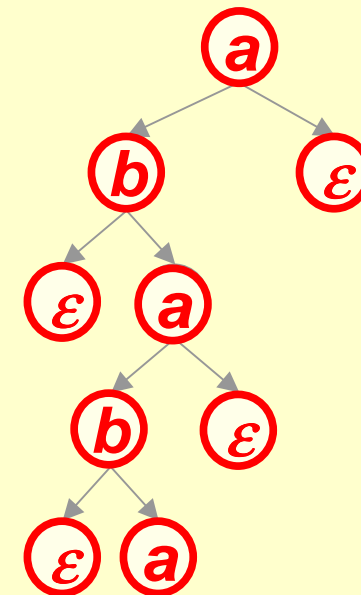
Iniciación AAAD

q	$p(q)$	n	$p(n/q)$
a	$21/50$	0	$13/21$
		2	$8/21$
b	$21/50$	0	$13/21$
		2	$8/21$
ϵ	$8/50$	0	$8/8$

preorden

```
send(a,2)
send(b,2)
send(ε,0)
send(a,2)
send(b,2)
send(ε,0)
send(a,0)
send(ε,0)
send(ε,0)
```

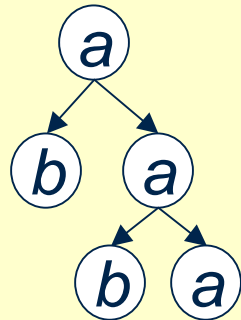
binario



Compresión mediante modelos k -testables adaptativos

Método 2: modelo contexto finito para árboles n -arios.

n -ario



Fase 1:

$a|b|b(ba)^*$

send_r(a)
send(a(ba))
send(b)
send(a(ba))
send(a(ab))
send(b)
send(a)

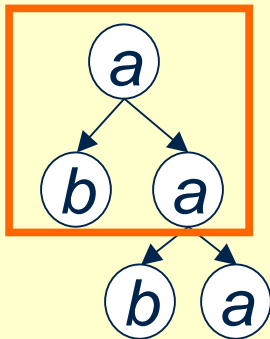
Fase 2:

send(a,2)
send(b,2)
send(ϵ ,0)
send(a,2)
send(b,2)
send(ϵ ,0)
send(a,0)
send(ϵ ,0)
send(ϵ ,0)

Compresión mediante modelos k -testables adaptativos

Método 3: predicción por concordancia parcial para árboles n -arios.

n -ario



modelo

1. encode_r(a(ba),3)
 encode_r(a,2)

 encode_p(a(ba),2)

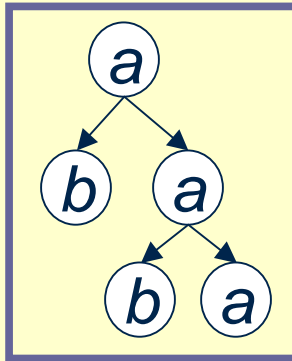
compresión

send_r(ϵ ,2)
send_r(ϵ ,2)
send_1(a)
send(ϵ ,a,2)
send_1M(2,a)
send_1(b)
send_1(a)

Compresión mediante modelos k -testables adaptativos

Método 3: predicción por concordancia parcial para árboles n -arios.

n -ario



modelo

2. encode_p(a(ba(ba)),3)
encode_p(b,2)
encode_p(a(ba),2)

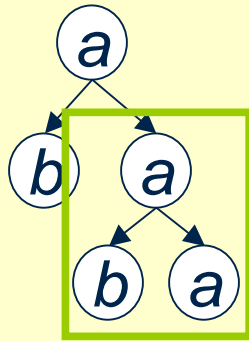
compresión

send(ϵ ,a(ba),3)
send(ϵ ,b,2)
send_1M(0,b)
send(a(ba),a(ba),2)

Compresión mediante modelos k -testables adaptativos

Método 3: predicción por concordancia parcial para árboles n -arios.

n -ario



modelo

3. encode_p(a(ba),3)
 encode_p(b,2)
 encode_p(a,2)

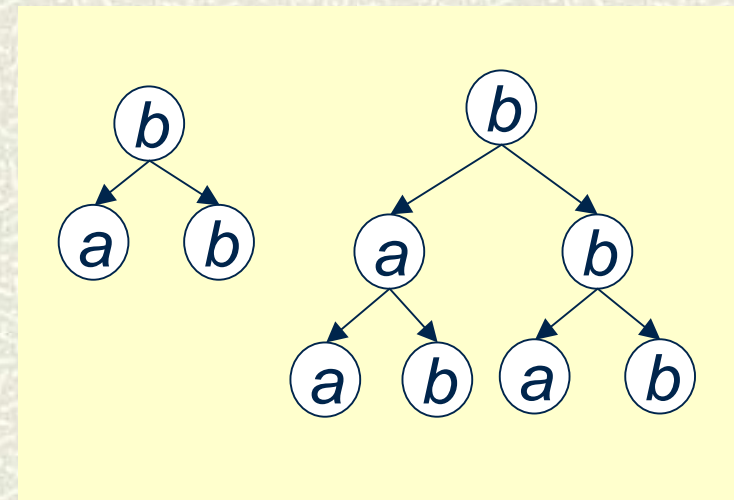
compresión

send(ϵ ,a(ba),3)
send(b,b,2)
send(ϵ ,a,2)
send_1M(0,a)

Compresión mediante modelos k -testables adaptativos

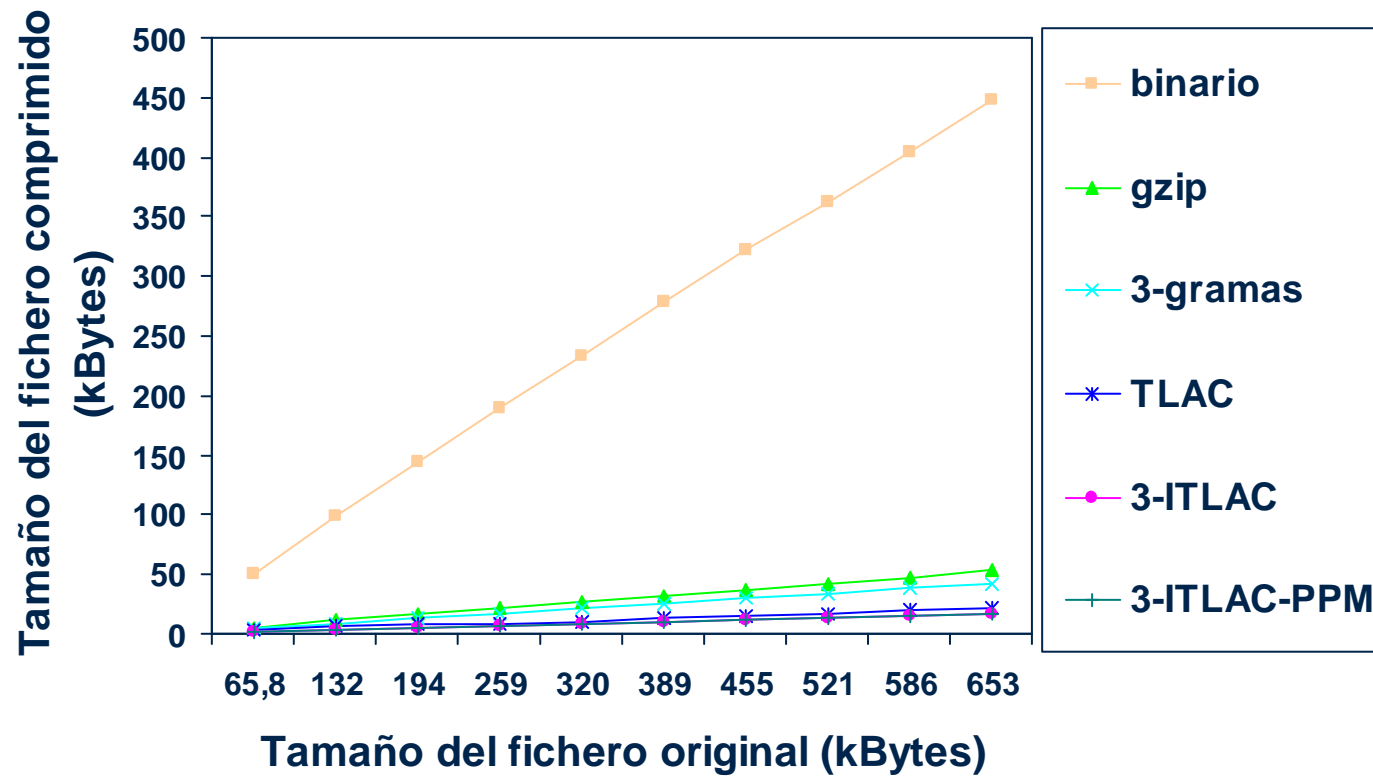
Modelo básico $M^{[1]}$

Compresión ($M=3$)			$M^{[1]}$
σ	$p(\sigma)$	m	$p_L(m/\sigma)$
a	$1/2$	0	$1/4$
		1	$1/4$
		2	$1/4$
		ε	$1/4$
b	$1/2$	0	$1/4$
		1	$1/4$
		2	$1/4$
		ε	$1/4$



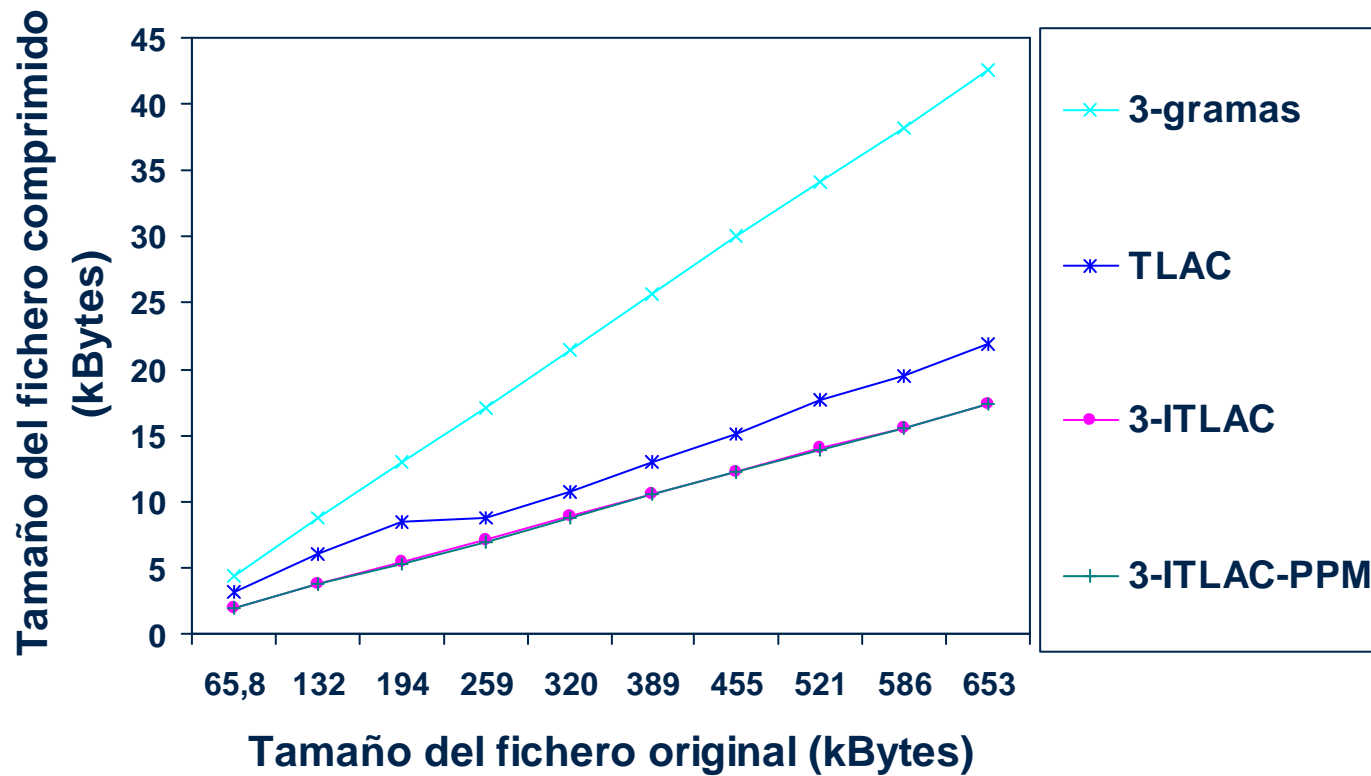
Compresión mediante modelos k -testables adaptativos

Resultados I



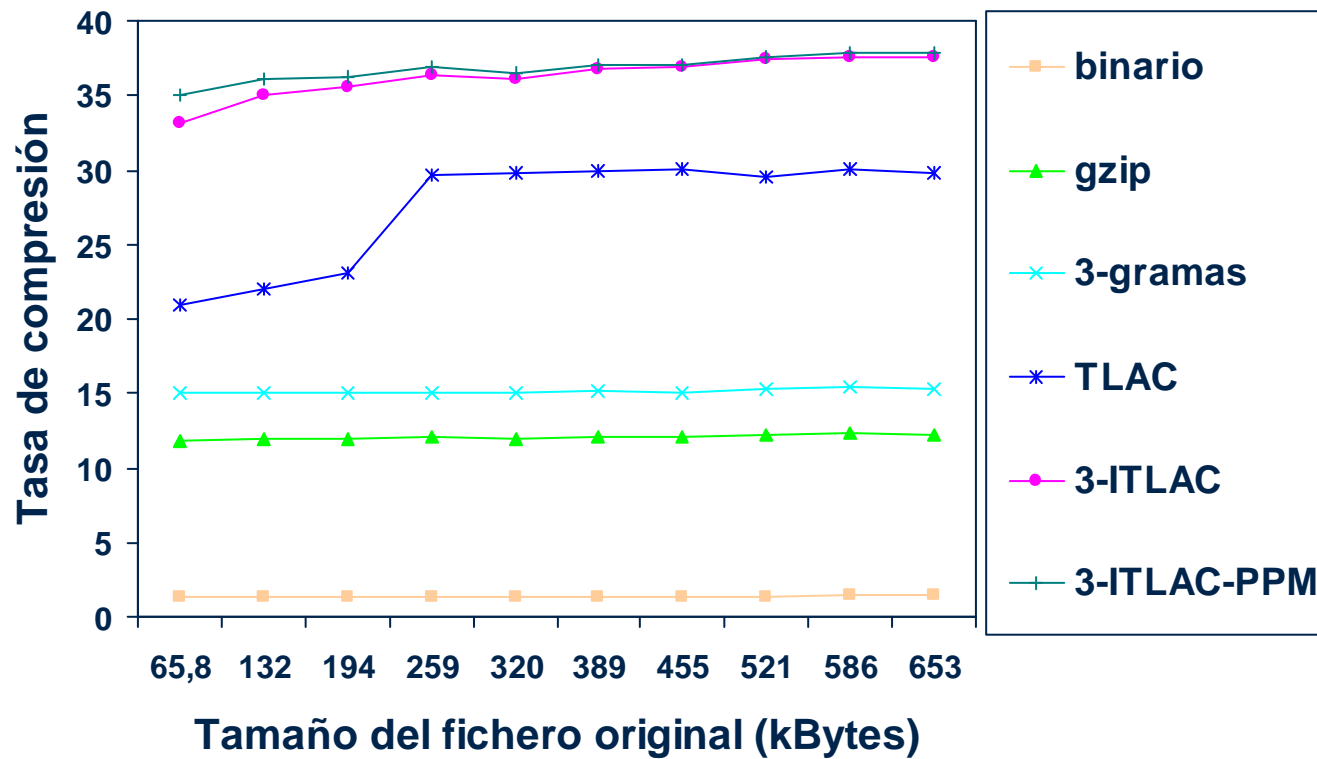
Compresión mediante modelos k -testables adaptativos

Resultados II



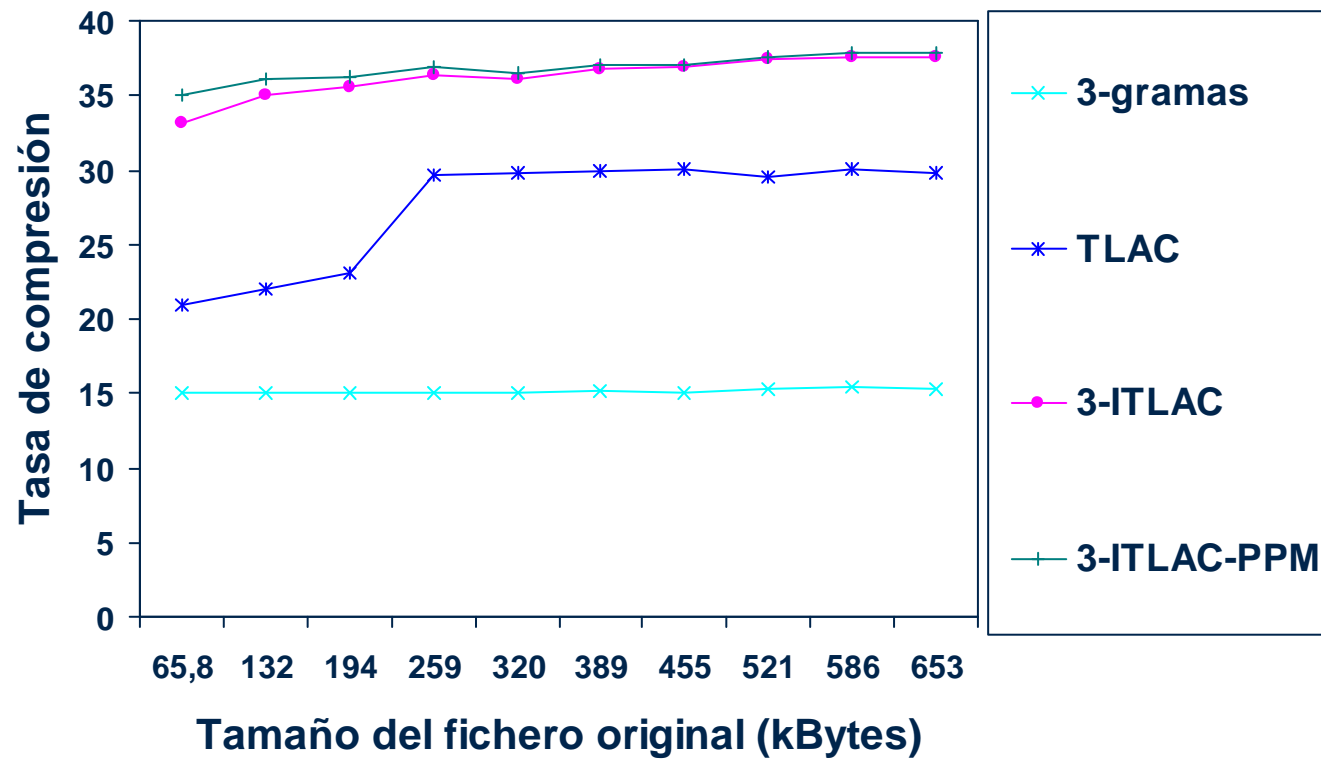
Compresión mediante modelos k -testables adaptativos

Resultados III



Compresión mediante modelos k -testables adaptativos

Resultados IV



Compresión mediante modelos k -testables adaptativos

Resultados V

Penn Tree-bank (tasa de compresión)	
gzip	6.52
3-gramas	9.08
3-ITLAC	9.10
bzip2	10.92
3-ITLAC-PPM	13.94

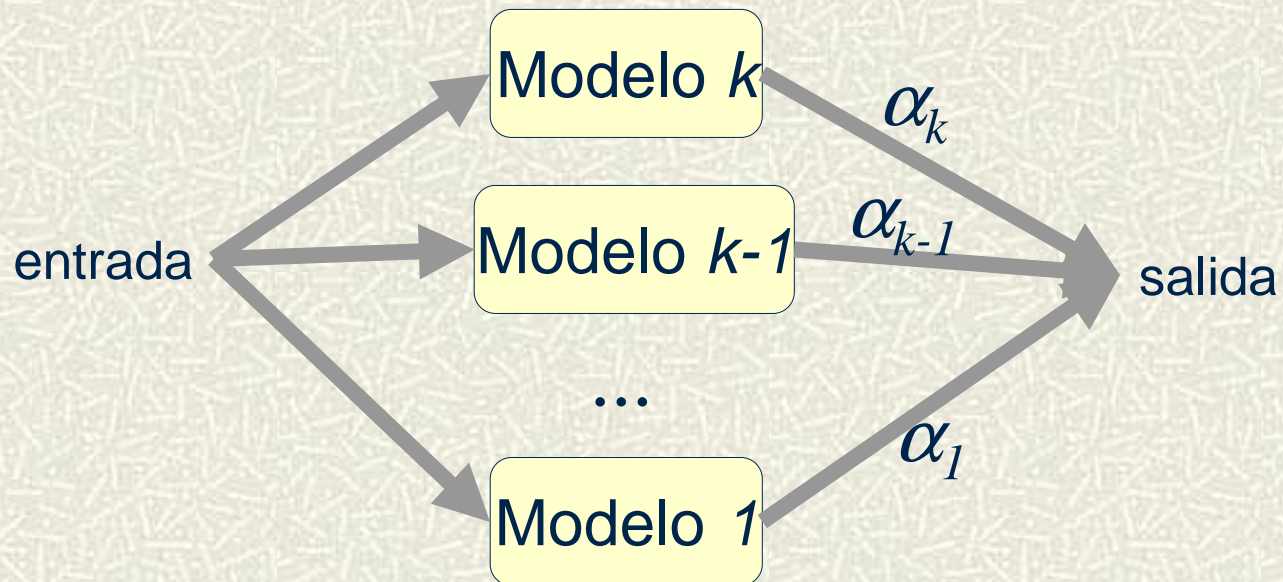
Compresión mediante modelos k -testables adaptativos

- Compresión de árboles binarios con inicialización a priori no mejora al Ziv-Lempel.
- Nuestros mejores resultados se obtienen con los modelos basados en PPM.
- Aplicación sobre datos:
 - *Artificiales: Mejora en 2.5 al gzip*
 - *Penn Tree-bank: Mejora en 2 al gzip*

Clasificación mediante modelos k -testables

Métodos de suavizado I:

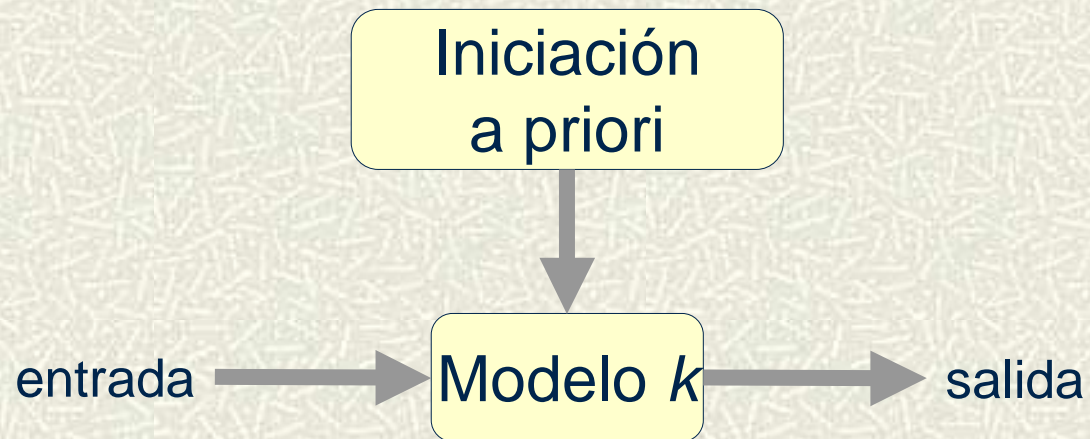
- Interpolación de modelos.



Clasificación mediante modelos k -testables

Métodos de suavizado II:

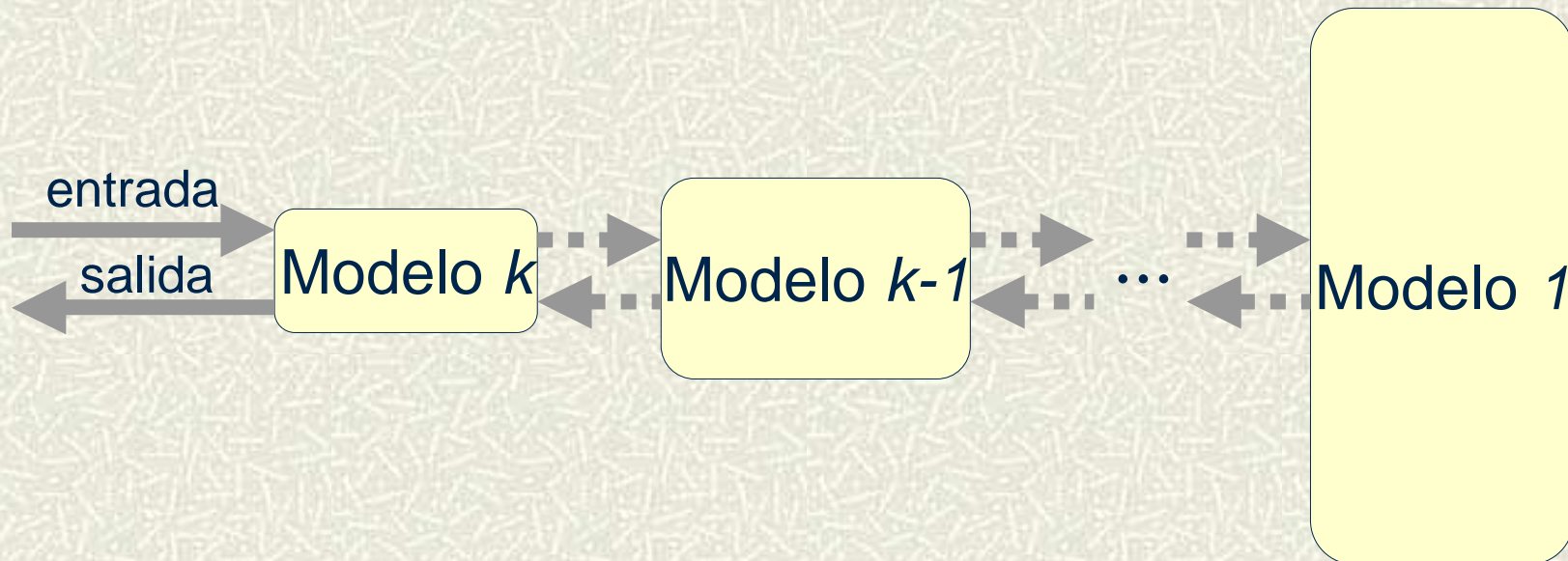
- Suavizado mediante distribución a priori.



Clasificación mediante modelos k -testables

Métodos de suavizado III:

- Predicción por concordancia parcial



Clasificación mediante modelos k -testables

- Predicción por concordancia parcial
Cálculo de la probabilidad

$$t = \sigma(t_1, \dots, t_m)$$

$$p(t|M) = p(t|M^{[kMax]})$$

$$p_m^{[k]}(\sigma, t_1, \dots, t_m) = fr(t)$$

Clasificación mediante modelos k -testables

- Predicción por concordancia parcial
Cálculo de la probabilidad

$$t = \sigma(t_1, \dots, t_m)$$

$$p(t|M) = p(t|M^{[kMax]})$$

$$p_m^{[k]}(\sigma, t_1, \dots, t_m) = \begin{cases} fr(t) - descuento(t) & \text{si } fr(t) > 0 \\ \frac{\sum descuento(t)}{\text{normalización}} \prod_{j=1}^m p_j^{[k-1]} & \text{en otro caso} \end{cases}$$

Clasificación mediante modelos k -testables

normalización

✘ Problema con la definición de factor de normalización

$$= \sum p(t|M^{[k-1]}) : fr(t)=0 \text{ en } M^{[k]}$$

✓ Solución: Cálculo del complementario

$$= 1 - \sum p(t|M^{[k-1]}) : fr(t)>0 \text{ en } M^{[k]}$$

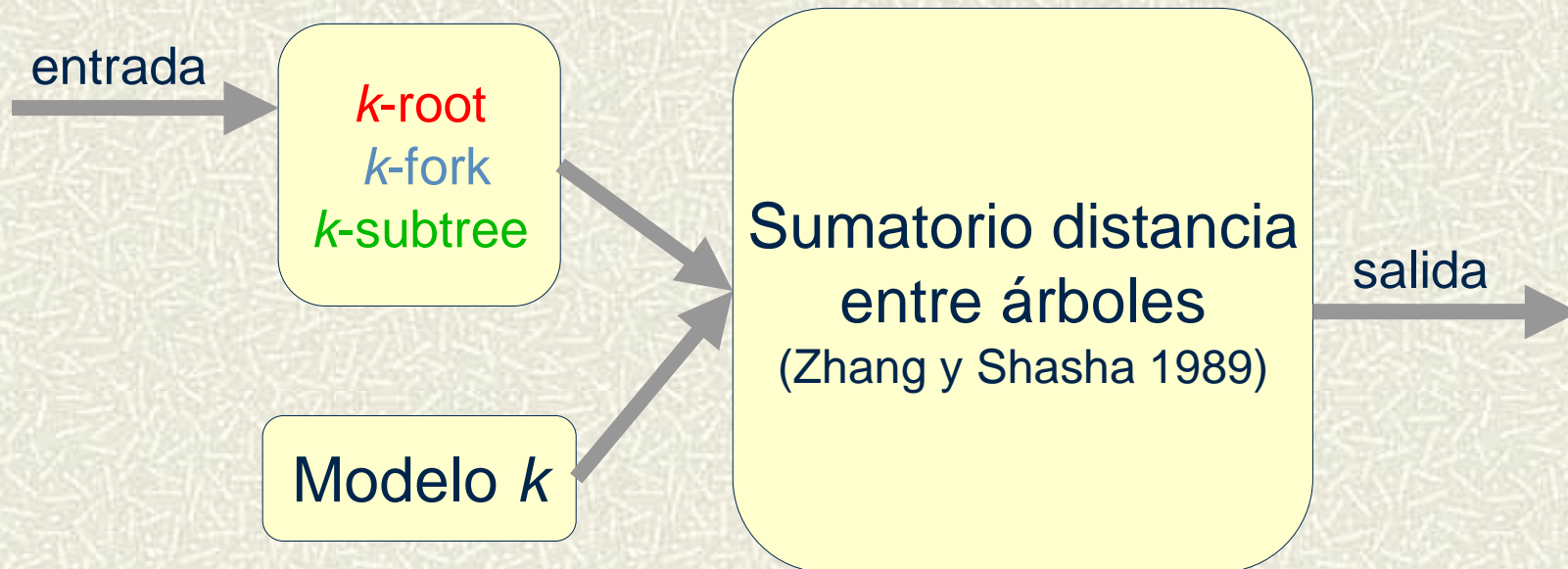
Clasificación mediante modelos k -testables

Modelo básico $M^{[1]}$

Clasificación $M^{[1]}$ $\mu=0.8$				
σ	$p(\sigma)$	μ_σ	m	$p_L(m \sigma)$
ε	$\Lambda_r=0$			$P(m 0.8)$
a	$4/10 (1-\Lambda_r)$	0.5	ε	$\Lambda_L(a) \frac{P(m 0.5)}{1-(P(0 0.5)+P(2 0.5))}$
			0	$3/4 (1-\lambda_L(0,a))$
			2	$1/4 (1-\lambda_L(2,a))$
b	$6/10 (1-\Lambda_r)$	1	ε	$\Lambda_L(b) \frac{P(m 1)}{1-(P(0 1)+P(2 1))}$
			0	$3/6 (1-\lambda_L(0,b))$
			2	$3/6 (1-\lambda_L(2,b))$

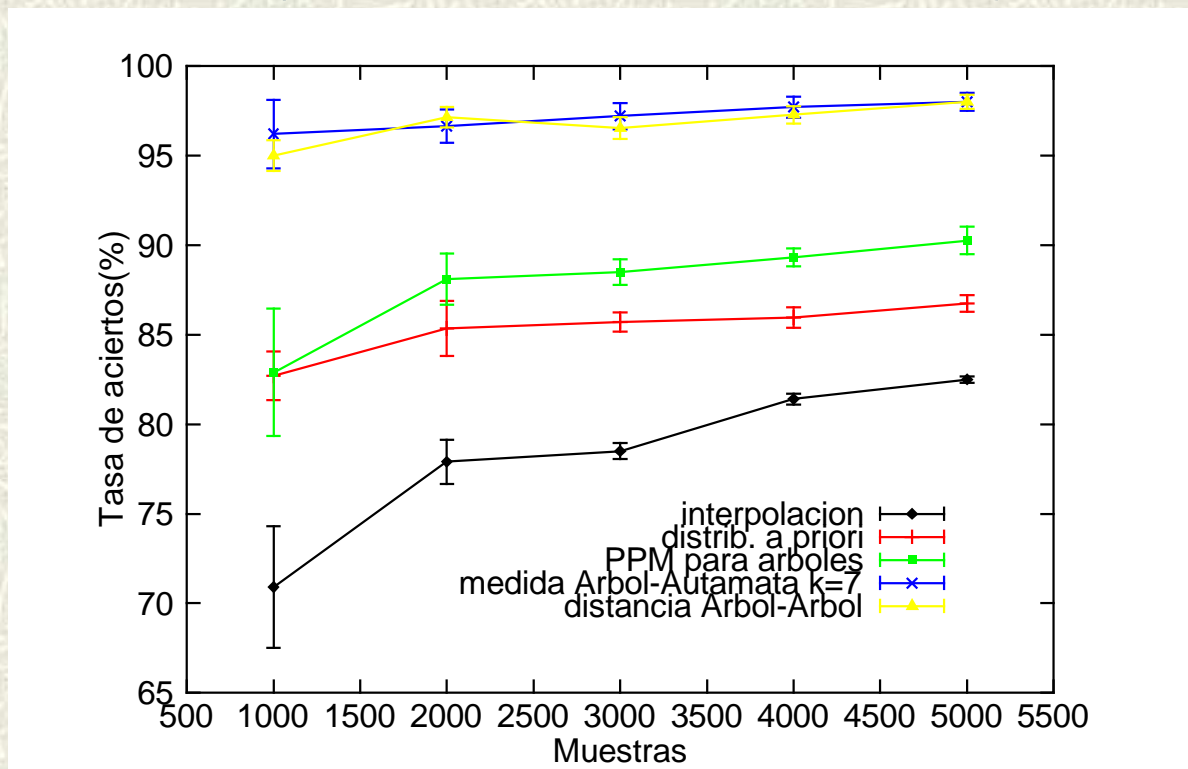
Clasificación mediante modelos k -testables

Clasificación no probabilística.



Clasificación mediante modelos k -testables

Resultados (NIST SPECIAL DATABASE 3).



Clasificación mediante modelos k -testables

- Hay dos tipos de modelos:
 - Probabilísticos (tasas [82%,90%]).
 - No probabilísticos (tasas [98%]).
- Los modelos no probabilísticos son extremadamente lentos para bases de datos extensas.
 $O(m \cdot n \cdot d \cdot n_b \cdot d_b)$ vs. $O(n \cdot \log m)$.

Parte II

Otros modelos de árboles

Compresión de superficies

Esquema de compresión

Árbol de expansión mínima (MST):

- Coordenadas relativas
- Distancia Euclídea

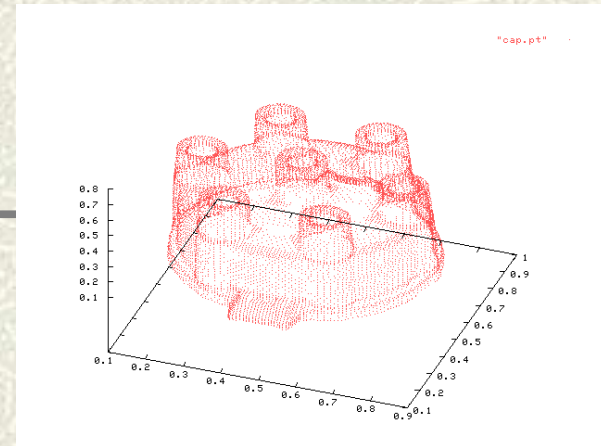
Modelo probabilístico:

- p_k : probabilidad de k hijos
- $F_i(x) = 1 / (1 + \exp(-\lambda_i(x - \mu_i)))$
- 3D: F_x , F_y y F_z

Codificación aritmética:

- Precisión ε : $F_i(x_t + \varepsilon / 2) - F_i(x_t - \varepsilon / 2)$
- Datos: $p_{k1} \mathbf{d}(n_1) p_{k2} \mathbf{d}(n_2) \dots p_{km} \mathbf{d}(n_m)$

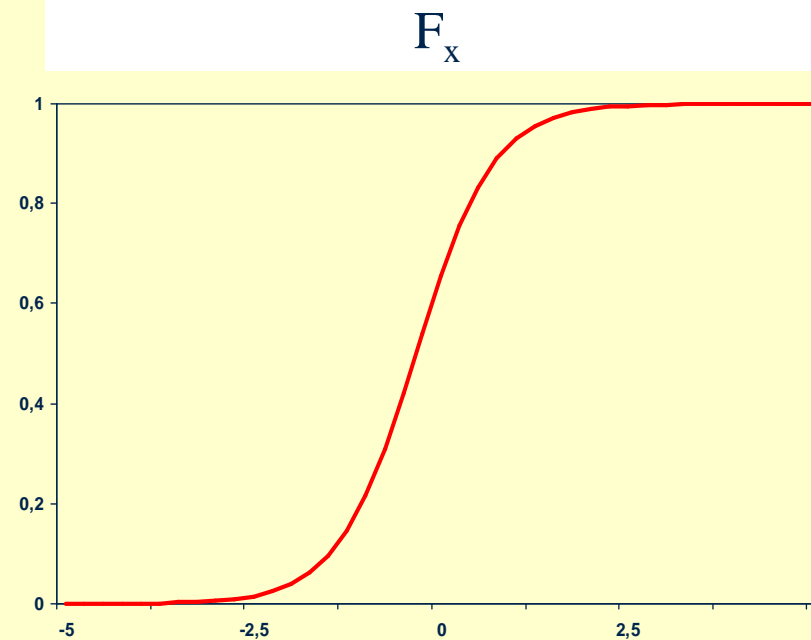
Datos comprimidos



Compresión de superficies

Ejemplo

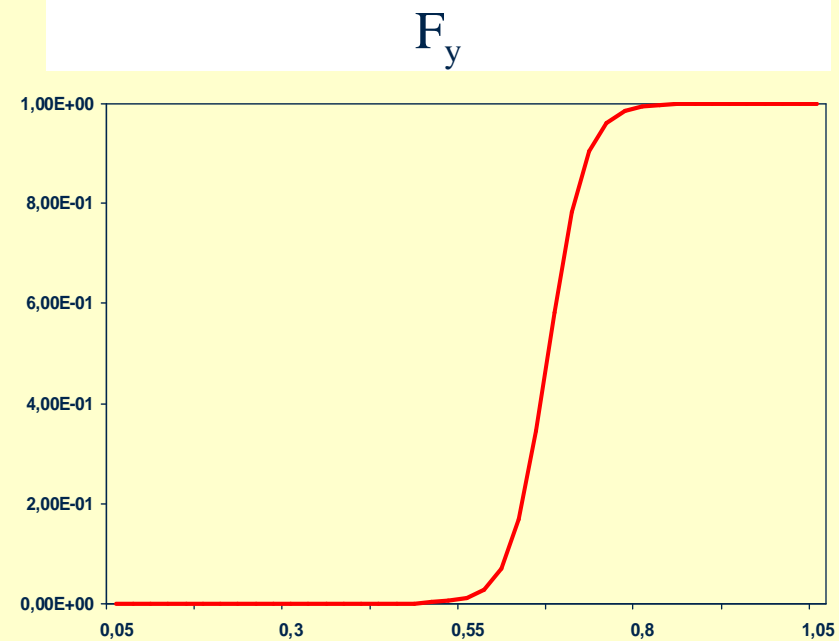
$$\lambda_x = 1.9238$$
$$\mu_x = -0.3333$$



Compresión de superficies

Ejemplo

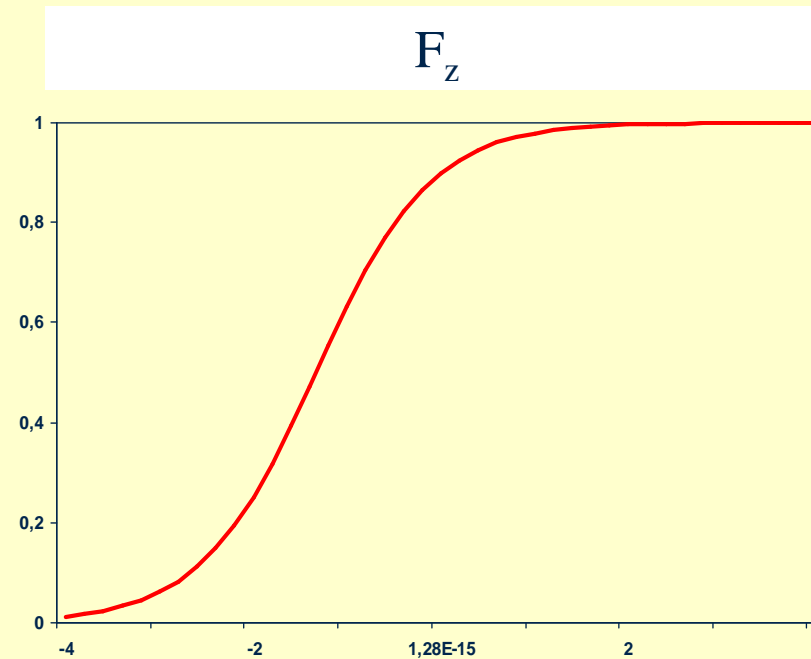
$$\lambda_y = 38.4765$$
$$\mu_y = 0.6666$$



Compresión de superficies

Ejemplo

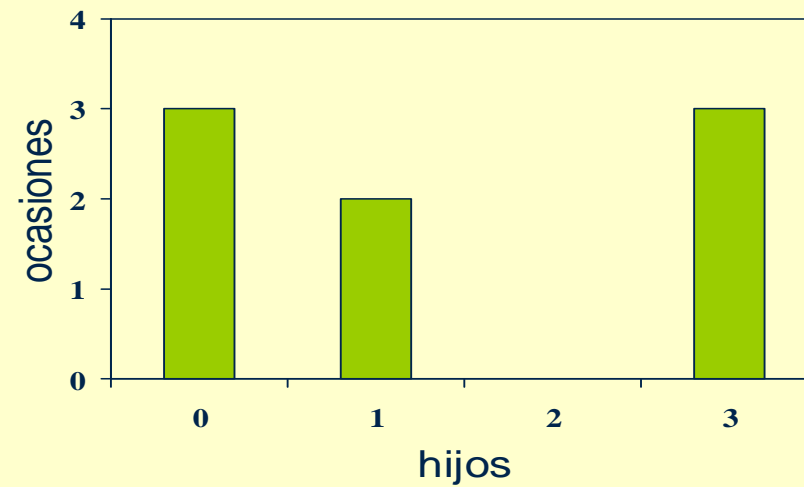
$$\lambda_z = 1.6407$$
$$\mu_z = -1.3333$$



Compresión de superficies

Ejemplo

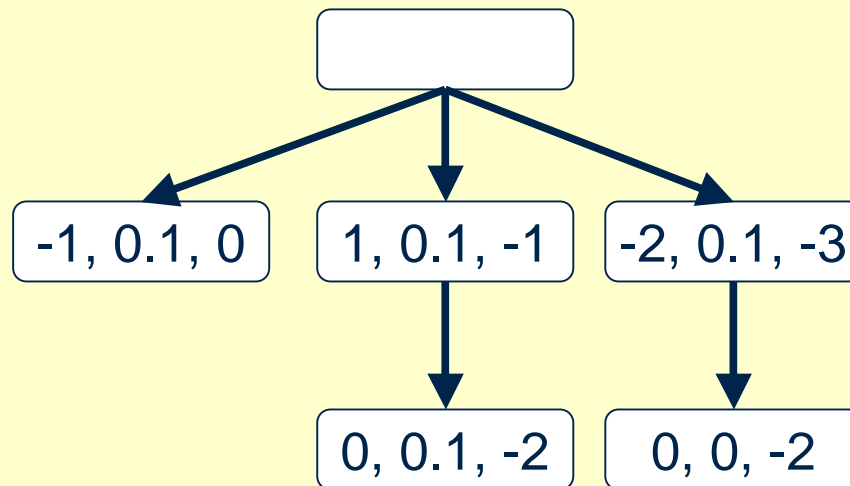
p_k : probabilidad k hijos



Compresión de superficies

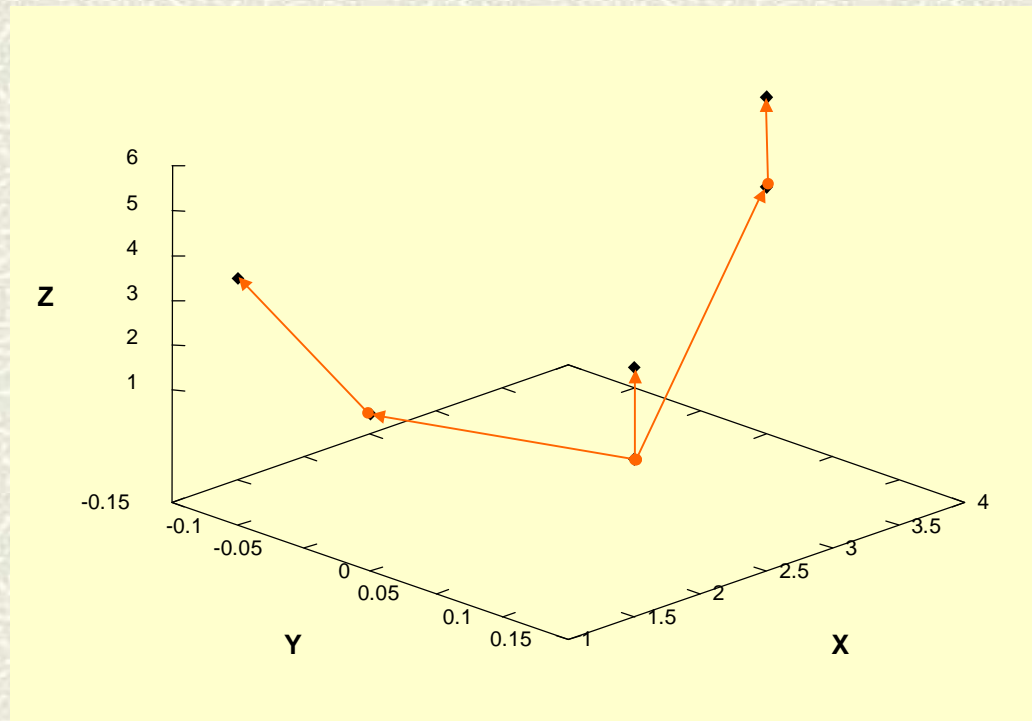
Ejemplo

Coordenadas relativas del MST



Compresión de superficies

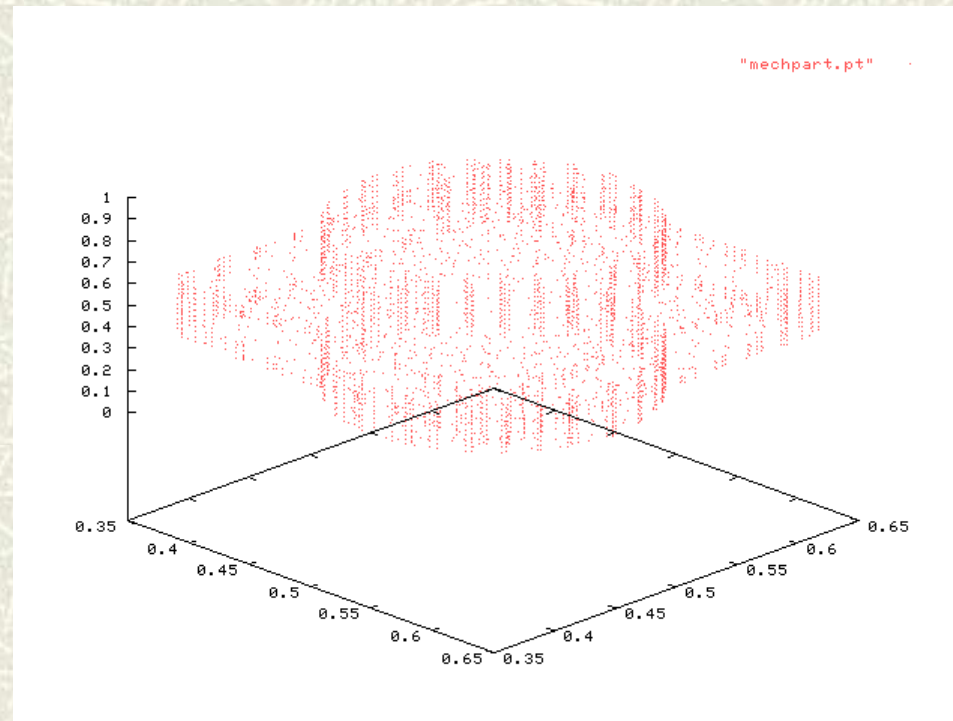
Ejemplo



X	Y	Z
2	0.1	1
3	0	1
1	-0.1	4
1	0	2
4	0	4
4	0	6

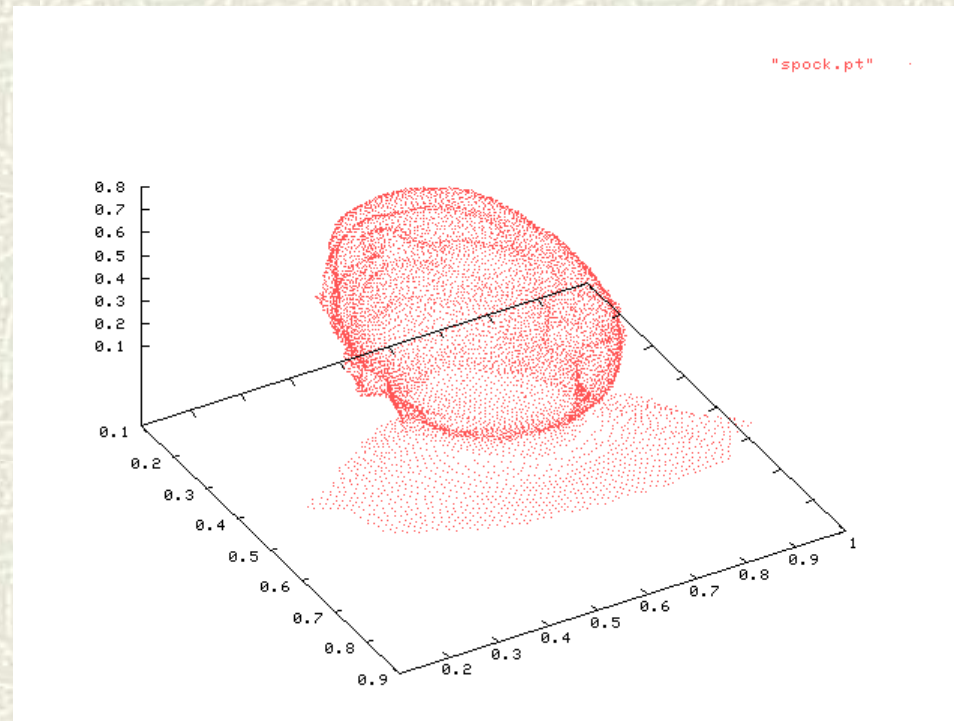
Compresión de superficies

Figuras



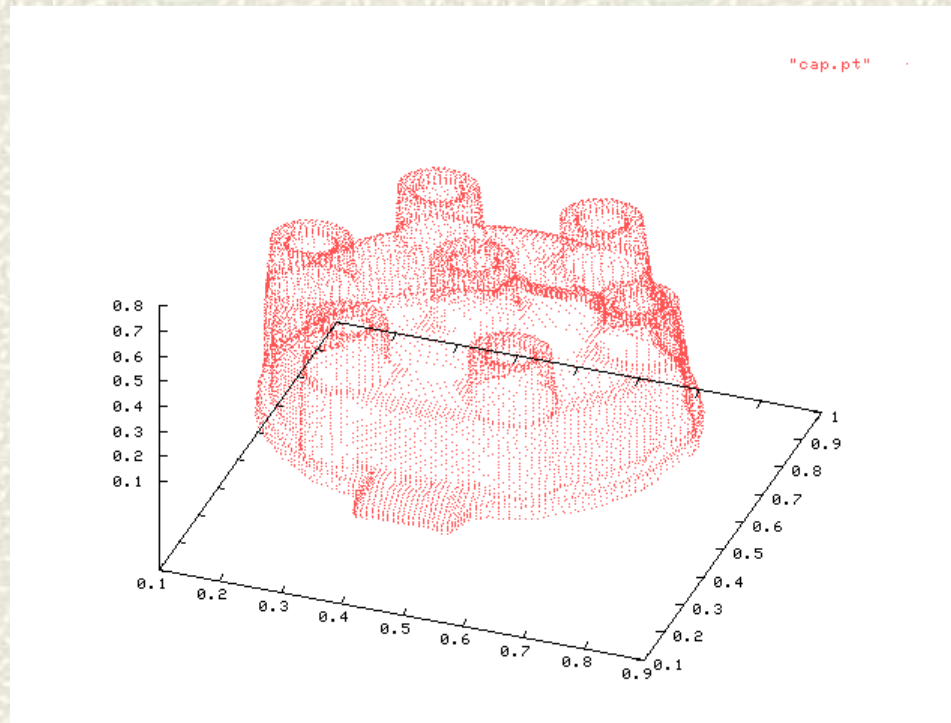
Compresión de superficies

Figuras



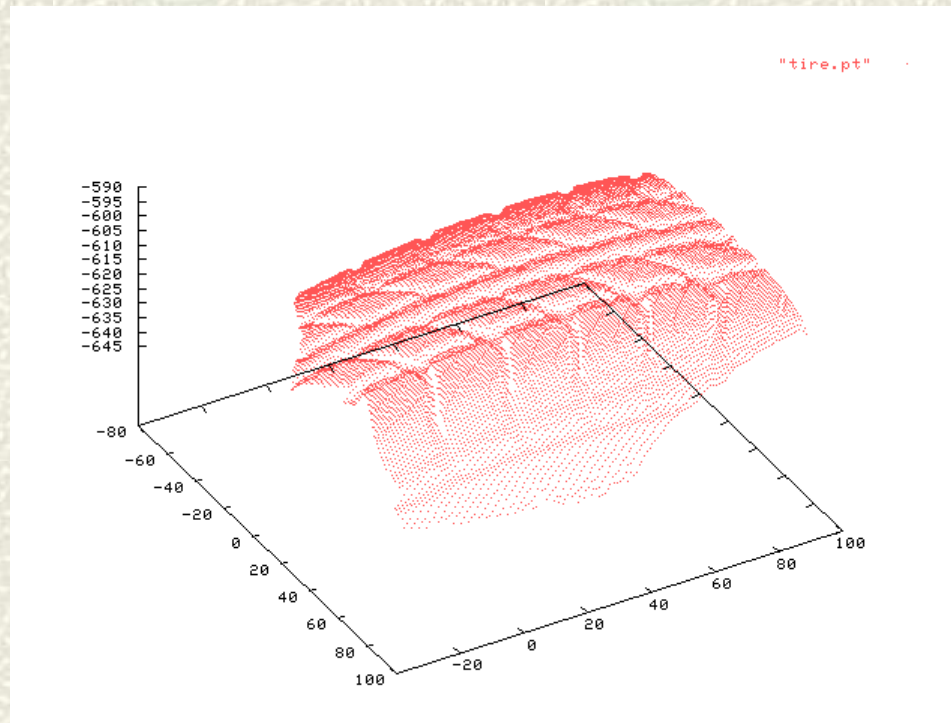
Compresión de superficies

Figuras



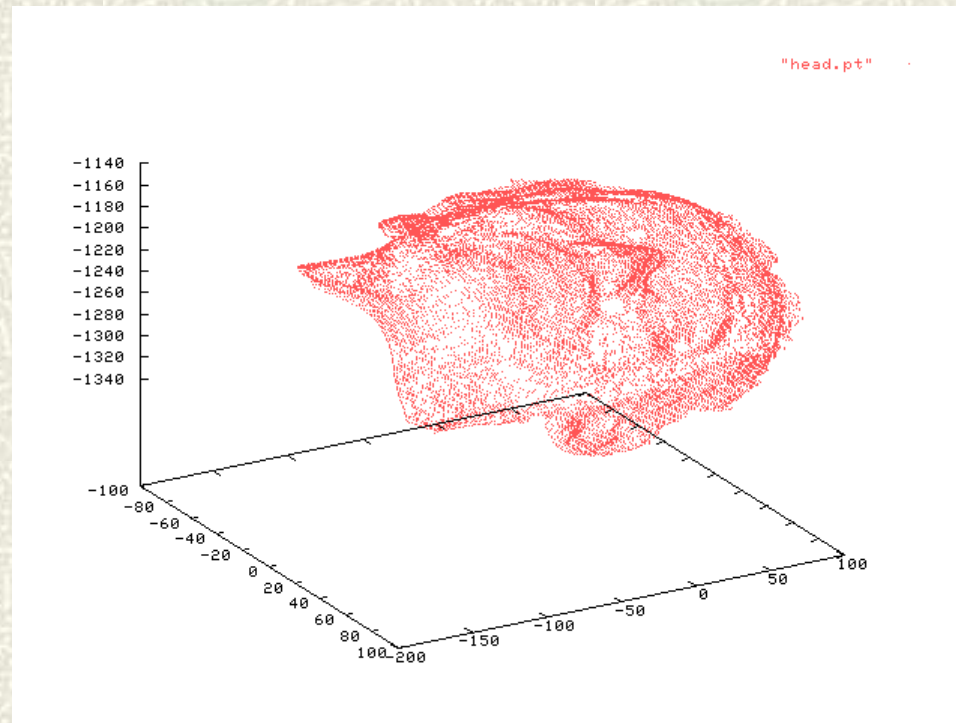
Compresión de superficies

Figuras



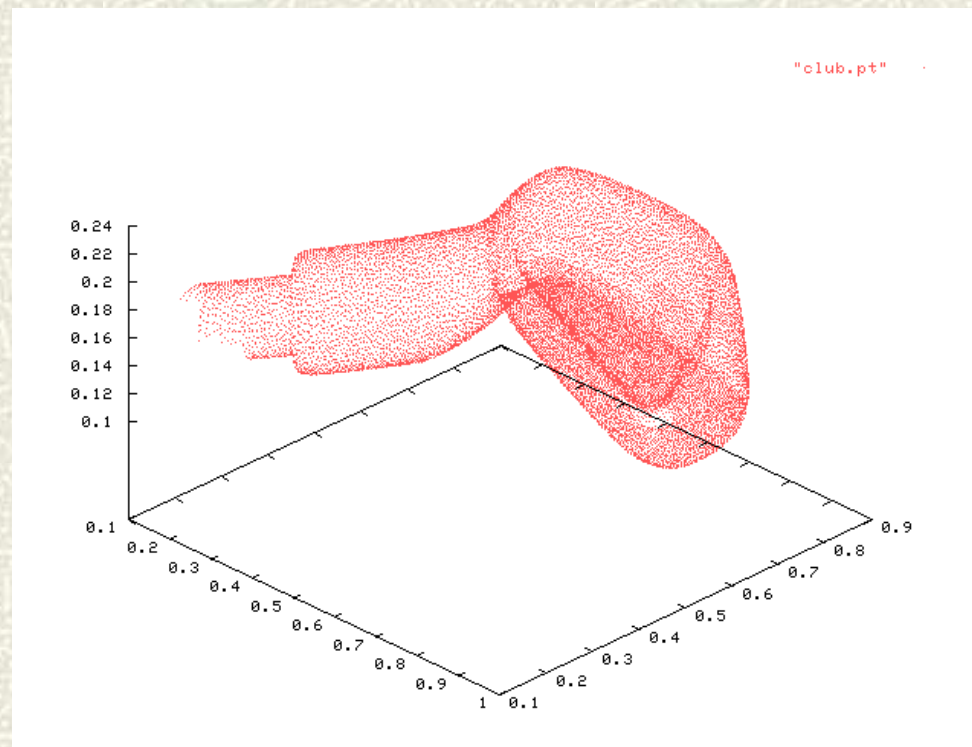
Compresión de superficies

Figuras



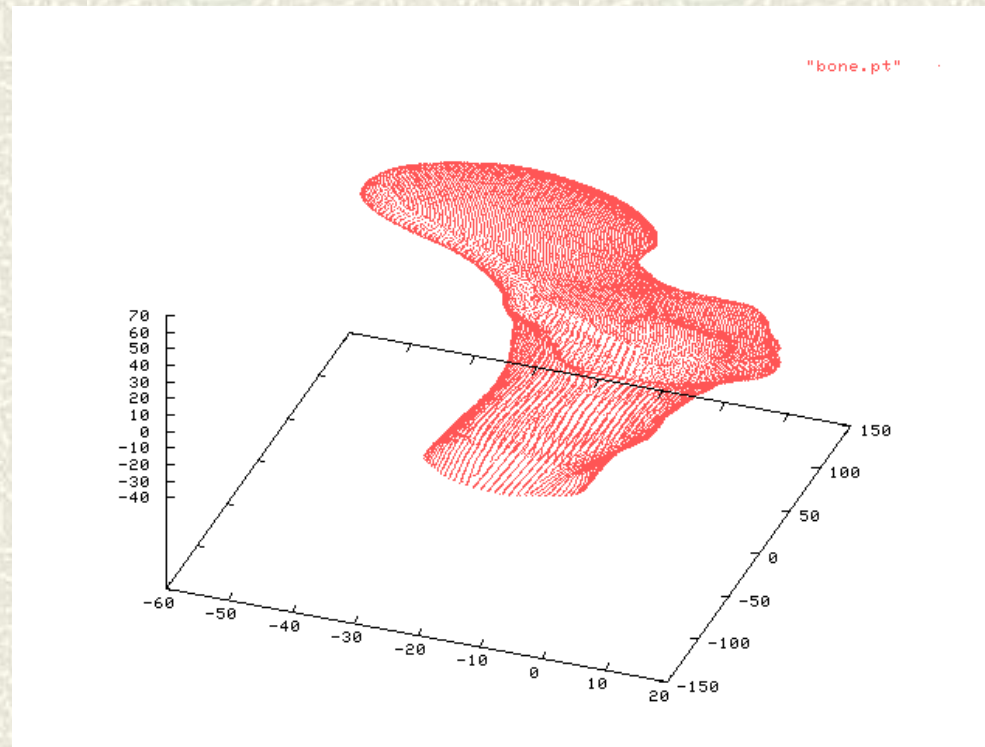
Compresión de superficies

Figuras



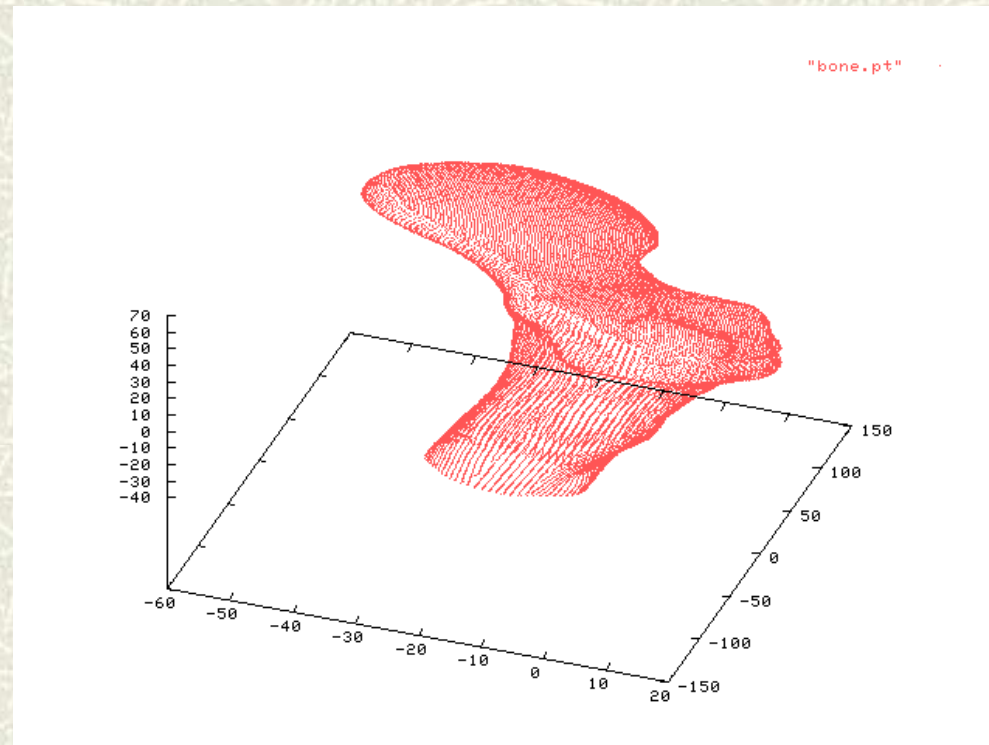
Compresión de superficies

Figuras



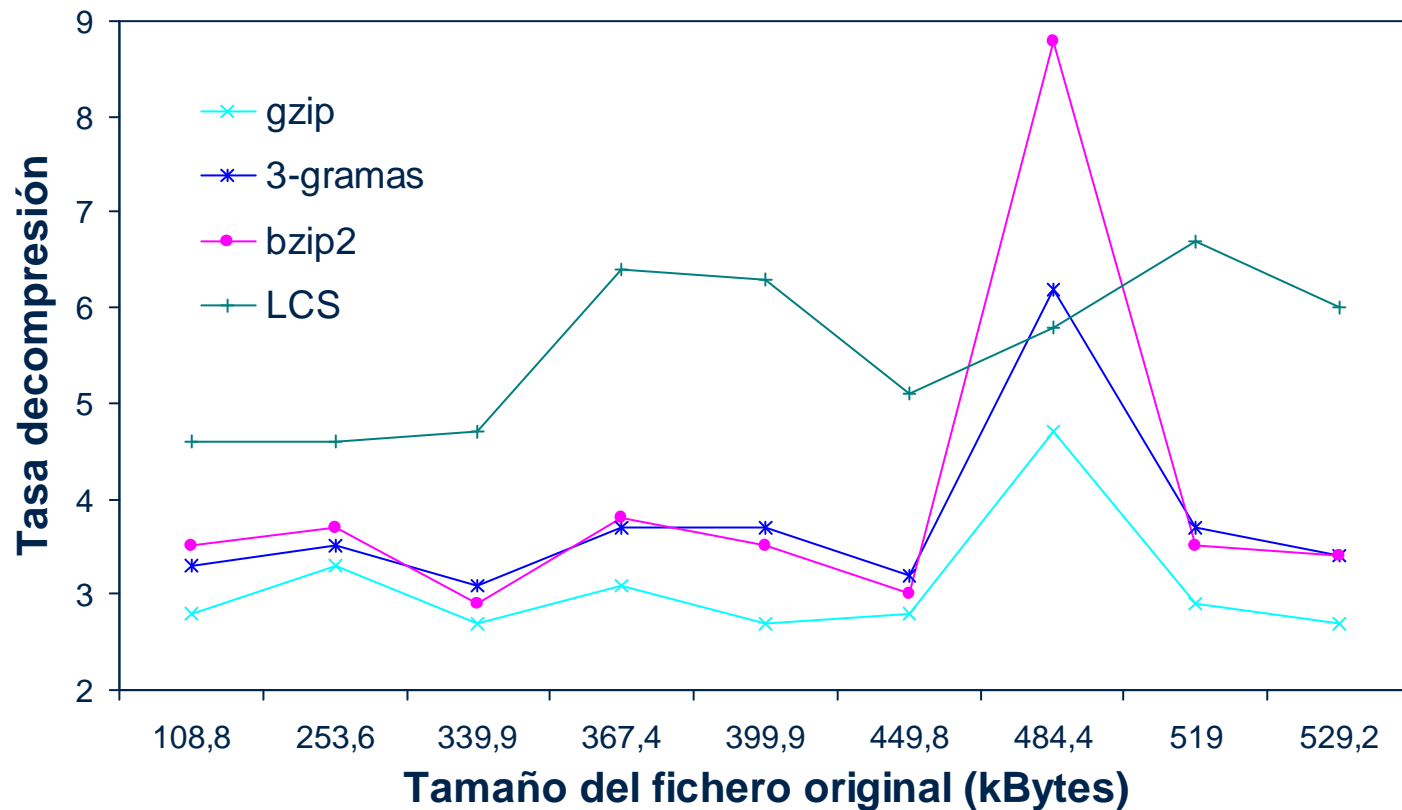
Compresión de superficies

Figuras



Compresión de superficies

Resultados y discusión



Compresión de superficies

- Codificador aritmético para superficies de puntos.
- Complejidad del método coincide con la del cálculo del MST.
- Mejores tasas que usando compresores de propósito general.
- Estudiar función 3D en vez de (F_x, F_y, F_z) .

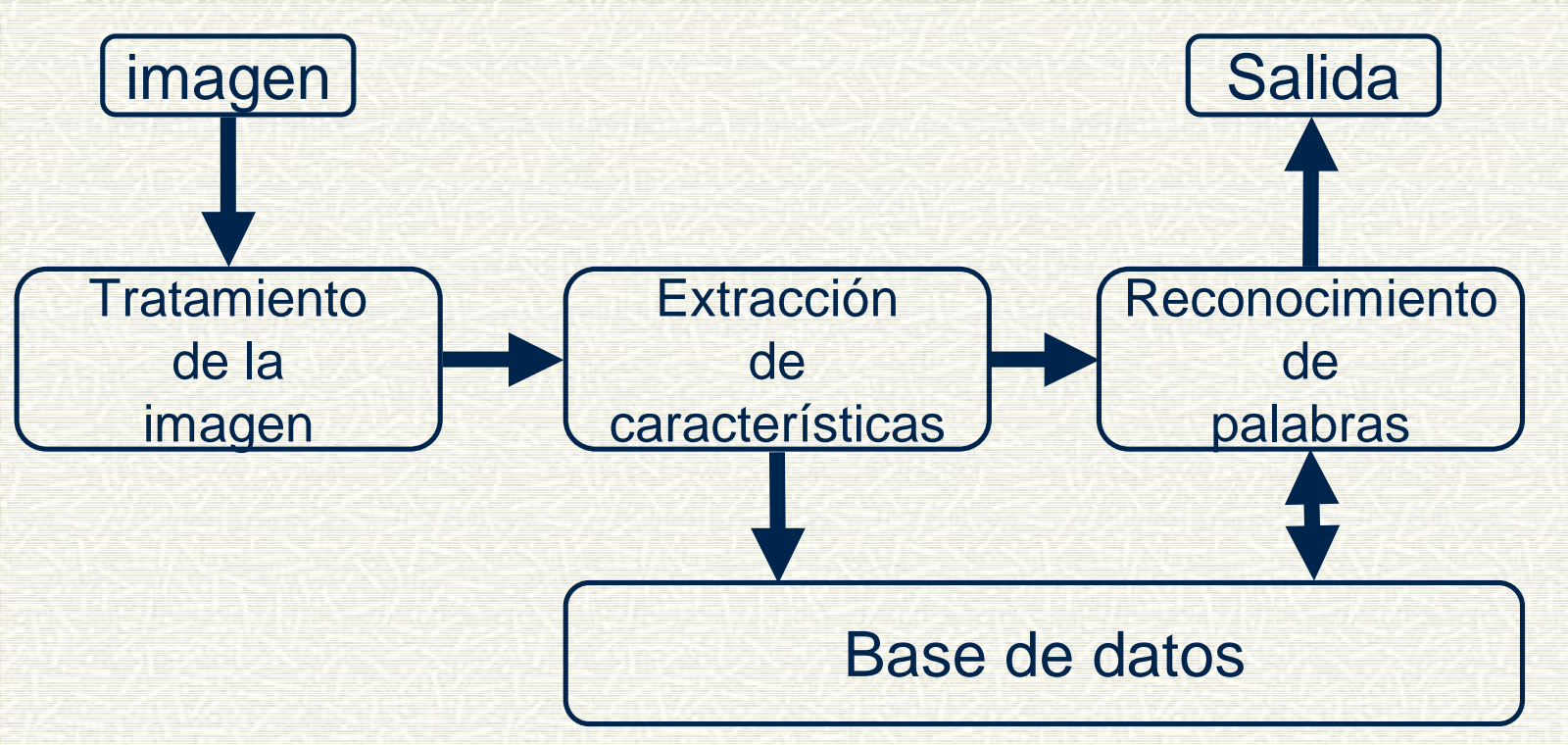
Reconocimiento de palabras manuscritas

Introducción

- Propósito: reconocer palabras completas.
- Letras aisladas (Suen et al. 1980; Elliman y Banks 1991; Bunke et al. 1995).
- Localización de palabras en un texto (Wang et al. 1997)

Reconocimiento de palabras manuscritas

Esquema general del reconocedor



Reconocimiento de palabras manuscritas

Imagen y características

- Tratamiento de la imagen

escala grises



B/N



Apertura morfológica
(Serra 1982)
+
esqueletizado
(Carrasco y Forcada 1995)

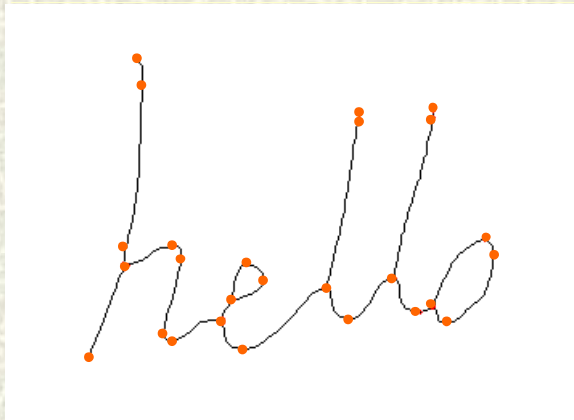


8 Reconocimiento de palabras manuscritas

Imagen y características

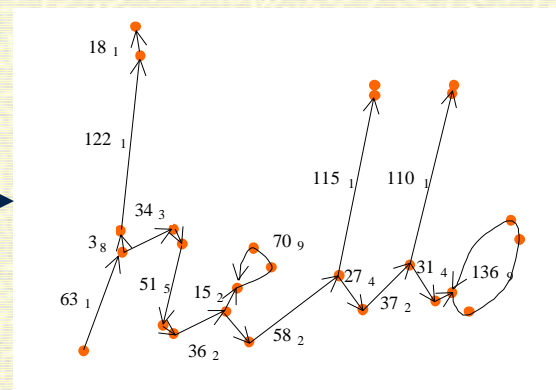
■ Extracción del árbol de características

Puntos dominantes
(Li y Yeaung 1997;
Powalka et al. 1997)



Cadena del árbol

Árbol
de
características



(63;1(34;3(9;4(51;5(8;4(36;2(15;2(70;9())28;4(58;2(115;1(4;1())27;
4(37;2(110;1(5;1())31;4(8;3(136;9())))))))))))3;8(122;1(18;1()))))

Reconocimiento de palabras manuscritas

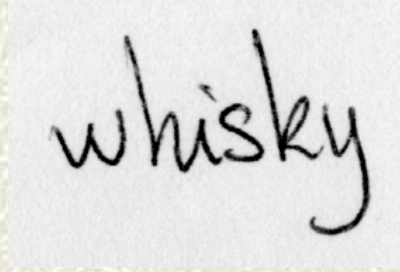
Clasificación

- Distancia de edición (Zhang y Shasha 1989)
- Técnica Leaving-One-Out (Duda y Hart 1973)
- Algoritmo adaptativo.

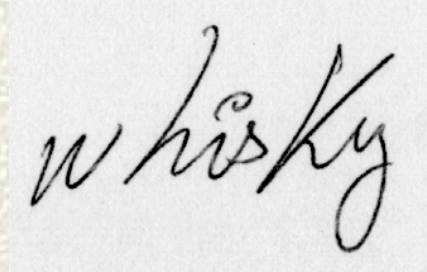
Reconocimiento de palabras manuscritas

Resultados

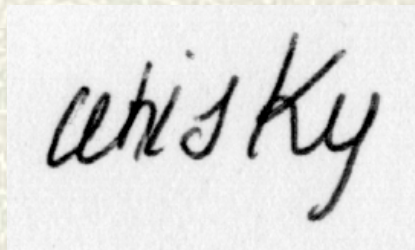
- Base datos:
 - #1: 50 palabras. 12 repeticiones. 4 escritores.



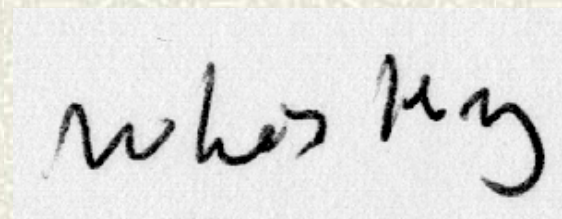
whisky



whisky



whisky

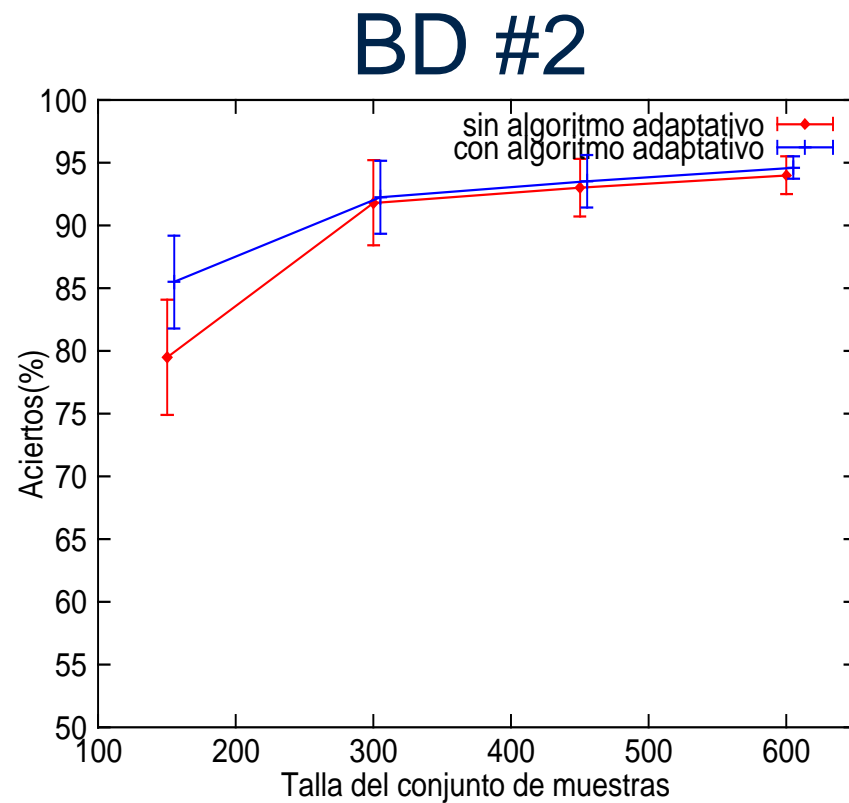
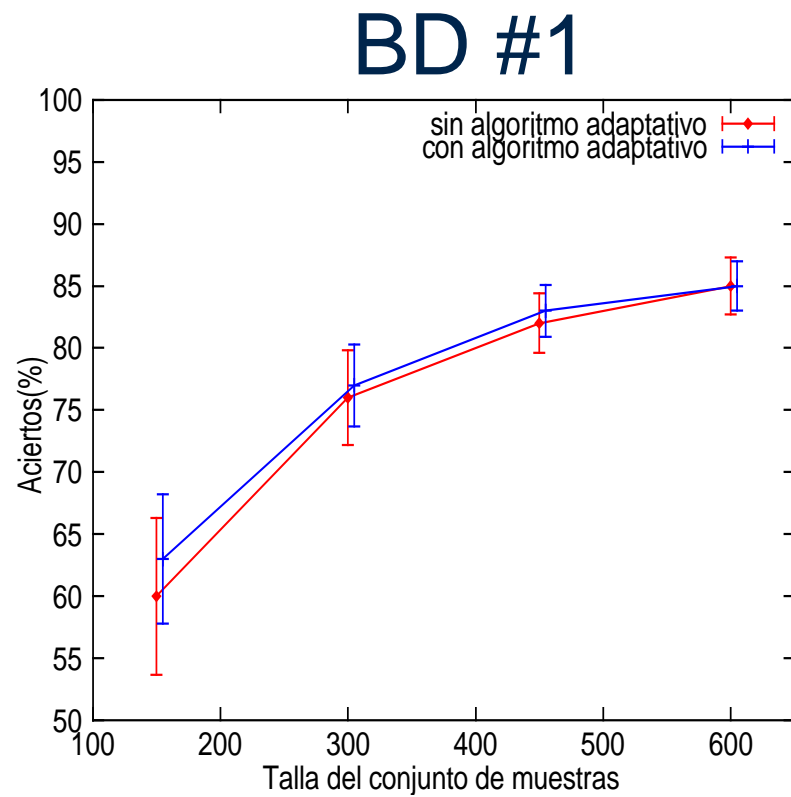


whisky

- #2: 34 palabras. 20 repeticiones. 1 escritor (LOB, Senior y Robinson 1998)

Reconocimiento de palabras manuscritas

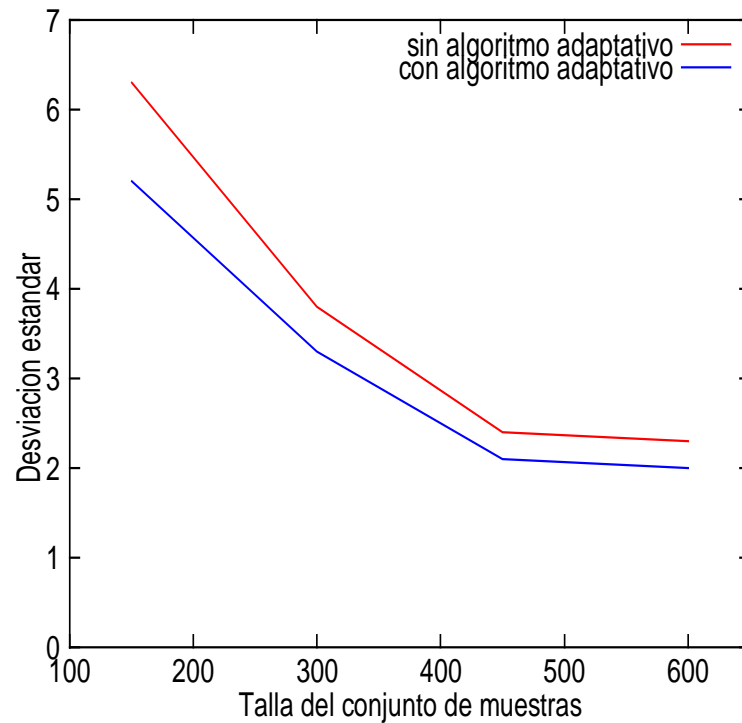
Resultados.



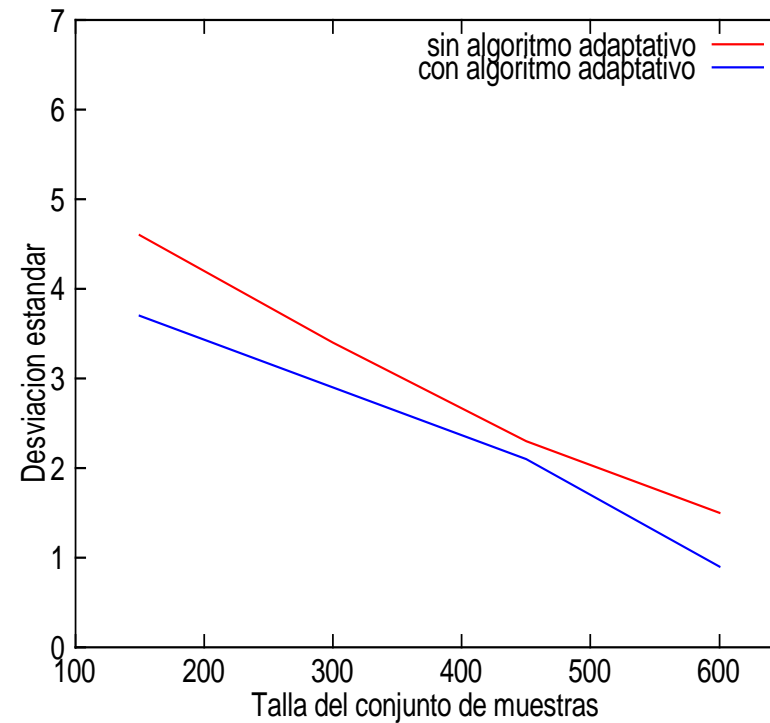
Reconocimiento de palabras manuscritas

Resultados.

BD #1



BD #2



Reconocimiento de palabras manuscritas

- Nuevo sistema de extracción de características.
- Algoritmo adaptativo.
- Futuros trabajos:
 - Coste de computación lineal con la BD. Se reduce con AESA (Micó et al. 1996; Micó y Oncina 1998).
 - Incrementar número de escritores.
 - Segmentación de palabras en letras aisladas.
 - Reconocimiento de firmas.

Parte III

Conclusiones y trabajos futuros

Conclusiones y trabajos futuros

Conclusiones

- Se ha definido una extensión de los lenguajes de árboles k -testables actualizable incrementalmente.
- Aplicación eficiente a la compresión y clasificación.
- Definición de un método de compresión para superficies 3D.
- Sistema de reconocimiento de palabras manuscritas basado en distancia de árboles

Conclusiones y trabajos futuros

Trabajos futuros

- Evaluación de bases de datos extensas. Especialmente lingüísticas para resolver ambigüedades léxicas y sintácticas.
- Generalizar los métodos de suavizado para gramáticas independientes del contexto.
- Aplicaciones al análisis de imágenes codificadas como árboles (bintree/quadtrees).
- Desarrollo de métodos eficientes para grafos dirigidos acíclicos.