



Selección de Objetos

Dr. Jesús Ariel Carrasco Ochoa

ariel@inaoep.mx

Oficina 8311



Contenido

- Introducción
- Estrategias de selección
- Técnicas filter
 - Condensación
 - Edición
- Técnicas wrapper



Selección de objetos

- Selección de instancias (Instance selection), Selección de prototipos (Prototype selection)
- Preprocesamiento
 - Clasificación Supervisada
 - Regresión



Selección de objetos

- Selección de instancias (Instance selection), Selección de prototipos (Prototype selection)
- Preprocesamiento
 - **Clasificación Supervisada**
 - Regresión



Selección de objetos

Por qué hacer selección de objetos en clasificación supervisada

- Mejorar los resultados de la clasificación
- Reducir el costo de la clasificación
- Acelerar el proceso de clasificación

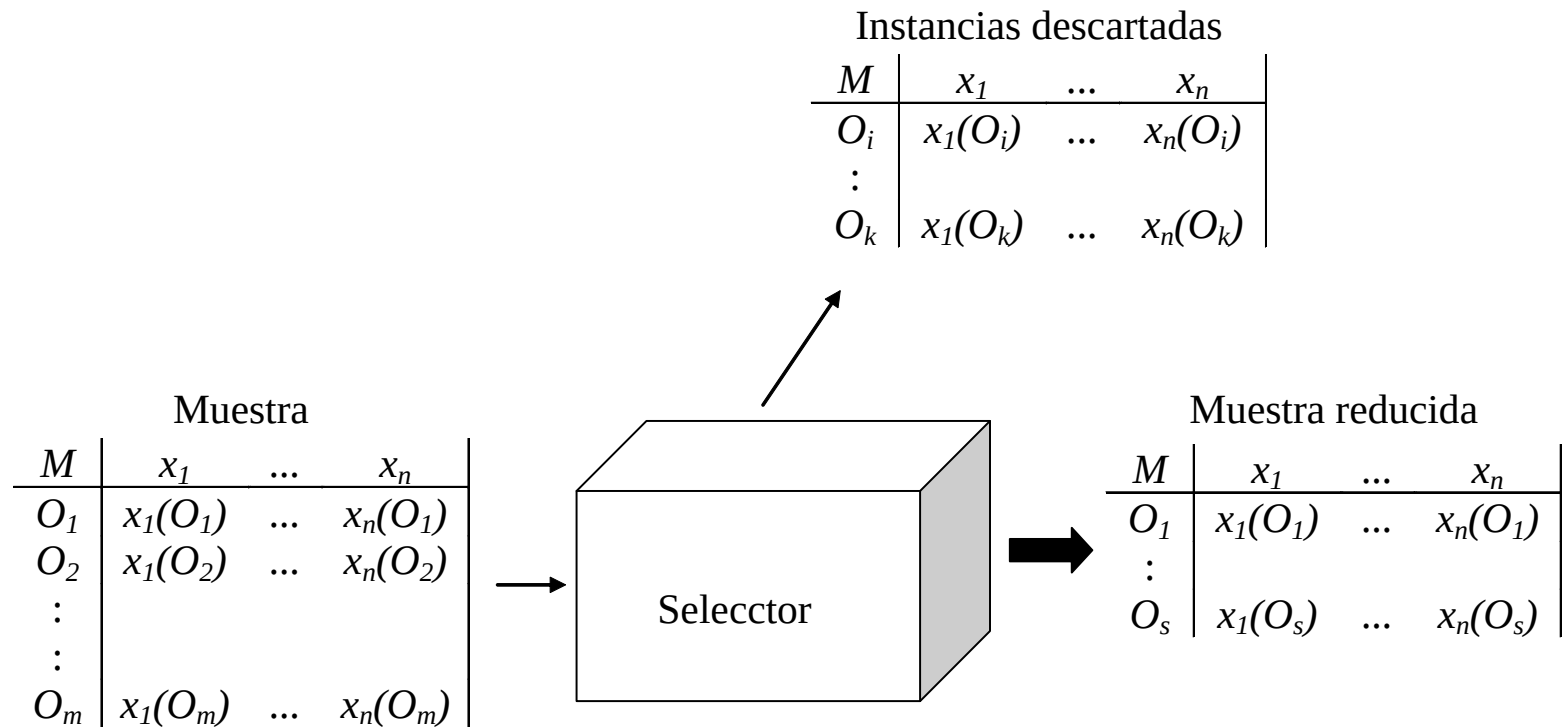


Selección de objetos

El objetivo es seleccionar un subconjunto reducido de objetos útiles para la clasificación eliminando

- Objetos irrelevantes
- Objetos redundantes
- Objetos ruidosos

Selección de objetos





Estrategias de Selección

- ***filter***.- La selección se hace con un criterio independiente del clasificador.
- ***wrapper***.- La selección se hace usando información del mecanismo de clasificación.



Estrategias filter

- Se busca eliminar diferentes tipos de objetos.
 - **Condensación:** Tratan de Eliminar objetos irrelevantes o redundantes
 - **Edición:** Tratan de eliminar outliers (ruido).
 - **Híbridos**



Métodos de Condensación

- Tratan de eliminar objetos que no son útiles para clasificar correctamente al conjunto de entrenamiento
- CNN (Hart, 1968)
- RNN (Gates, 1972)
- CNN+RNN (Gates, 1972)

CNN (Hart, 1968)

Dada una muestra D de objetos descritos por n atributos.

1. Inicializar en $S = \emptyset$
2. Para cada objeto $O \in D$
 3. Si el vecino más cercano a O en S es de clase diferente
 4. $S = S \cup \{O\}$
 5. $D = D - \{O\}$
6. Si se agregaron objetos a S ir a 2
7. Tomar $D = S$

RNN (Gates, 1972)

Dada una muestra D de objetos descritos por n atributos.

1. Inicializar en $S=D$
2. Para cada objeto $O \in S$
 3. Si $\forall O' \in D$ el vecino más cercano a O' en $S - \{O\}$ es de su misma clase
 4. $S = S - \{O\}$
5. Tomar $D=S$

CNN+RNN (Gates, 1972)

Dada una muestra D de objetos descritos por n atributos.

1. Inicializar en $S = \emptyset$
2. Para cada objeto $O \in D$
 3. Si el vecino más cercano a O en S es de clase diferente
 4. $S = S \cup \{O\}$
 5. $D = D - \{O\}$
6. Si se agregaron objetos a S ir a 2
7. Para cada objeto $O \in S$
 8. Si $\forall O' \in D$ el vecino más cercano a O' en $S - \{O\}$ es de su misma clase
 9. $S = S - \{O\}$
10. Tomar $D = S$



Métodos de Edición

- Tratan de eliminar objetos que no son bien clasificados o causan la mala clasificación de otros
- ENN (Wilson, 1972)
- RENN (Tomek, 1976)
- MultiEdit (Kittler, 1980)

ENN (Wilson, 1972)

Dada una muestra D de objetos descritos por n atributos.

1. Inicializar en $S=D$
2. Para cada objeto $O \in S$
 3. Si la mayoría de sus k vecinos más cercanos en D son de diferente clase
 4. $S = S - \{O\}$
5. $D=S$

RENN (Tomek, 1976)

Dada una muestra D de objetos descritos por n atributos.

1. Inicializar en $S=D$
2. Para cada objeto $O \in S$
 3. Si la mayoría de sus k vecinos más cercanos en S son de diferente clase
 4. $S = S - \{O\}$
5. Tomar $D=S$
6. Si se eliminó algún objeto de S ir a 1

MultiEdit (Kittler, 1980)

Dada una muestra D de objetos descritos por n atributos.

1. Hacer una partición aleatoria (D_0, D_1, D_r) de D tal que todos los D_i sean de aproximadamente el mismo tamaño.
2. Para cada D_i aplicar ENN pero buscando los vecinos en $D_{(i+1) \bmod r}$
3. Tomar $D = S_0 \cup S_1 \cup \dots \cup S_r$
4. Si se eliminó algún objeto ir a 1



Métodos de Híbridos

- Combinan ideas de Condensación y Edición.
 - Tratan de eliminar objetos que no son útiles para clasificar correctamente al conjunto de entrenamiento
 - Tratan de eliminar objetos que no son bien clasificados o causan la mala clasificación de otros
- DROP 1-5 (Wilson & Martínez, 2000)
- ICF (Brighton & Mellish, 1999, 2002)

DROP 1-5 (Wilson & Martínez, 2000)

Decremental Reduction Optimization Procedure.

1. Inicializar en $S=D$
2. Para cada objeto $O \in S$
 3. Si la mayoría de sus k vecinos más cercanos en D son de diferente clase
 4. $S = S - \{O\}$
5. $D=S$

DROP 1 (Wilson & Martínez, 2000)

Dada una muestra D de objetos descritos por n atributos.

1. Inicializar en $S=D$
2. Para cada objeto $O \in S$
 3. Si $\forall O' \in S$ el vecino más cercano a O' en $S - \{O\}$ es de su misma clase
 4. $S = S - \{O\}$
5. Tomar $D=S$

DROP 2 (Wilson & Martínez, 2000)

Dada una muestra D de objetos descritos por n atributos.

1. Inicializar en $S=D$
2. Ordenar S descendientemente de acuerdo a la distancia con su vecino más cercano de diferente clase
3. Para cada objeto $O \in S$
 4. Si $\forall O' \in D$ el vecino más cercano a O' en $S - \{O\}$ es de su misma clase
 5. $S = S - \{O\}$
6. Tomar $D=S$

DROP 3 (Wilson & Martínez, 2000)

Dada una muestra D de objetos descritos por n atributos.

1. Aplicar ENN
2. Inicializar en $S=D$
3. Ordenar S descendientemente de acuerdo a la distancia con su vecino más cercano de diferente clase
4. Para cada objeto $O \in S$
 5. Si $\forall O' \in D$ el vecino más cercano a O' en $S - \{O\}$ es de su misma clase
 6. $S = S - \{O\}$
7. Tomar $D=S$

DROP 4 (Wilson & Martínez, 2000)

Dada una muestra D de objetos descritos por n atributos.

1. Inicializar en $S=D$
2. Para cada objeto $O \in S$
 3. Si la mayoría de sus k vecinos más cercanos en D son de diferente clase y $\forall O' \in D$ el vecino más cercano a O' en $S - \{O\}$ es de su misma clase
4. $S = S - \{O\}$

DROP 4 (Wilson & Martínez, 2000)

5. Ordenar S descendientemente de acuerdo a la distancia con su vecino más cercano de diferente clase
6. Para cada objeto $O \in S$
 7. Si $\forall O' \in D$ el vecino más cercano a O' en $S - \{O\}$ es de su misma clase
 8. $S = S - \{O\}$
9. Tomar $D = S$

DROP 5 (Wilson & Martínez, 2000)

Dada una muestra D de objetos descritos por n atributos.

1. Inicializar en $S=D$
2. Ordenar S Ascendentemente de acuerdo a la distancia con su vecino más cercano de diferente clase
3. Para cada objeto $O \in S$
 4. Si $\forall O' \in D$ el vecino más cercano a O' en $S - \{O\}$ es de su misma clase
 5. $S = S - \{O\}$
6. Tomar $D=S$

ICF (Brighton & Mellish, 1999, 2002)

Dada una muestra D de objetos descritos por n atributos.

1. Aplicar ENN
2. Inicializar en $S=D$
3. Para cada objeto $O \in S$
 4. $\text{Reachable}(O)$ = número de objetos de su clase que son más cercanos que el objeto más cercano de otra clase.
 5. $\text{Coverage}(O)$ = número de objetos O' de su clase para los cuales O es más cercano a O' que el más cercano a O' de diferente clase
6. Para cada objeto $O \in S$
 7. Si $\text{Reachable}(O) > \text{Coverage}(O)$
 8. $S = S - \{O\}$
9. Tomar $D=S$

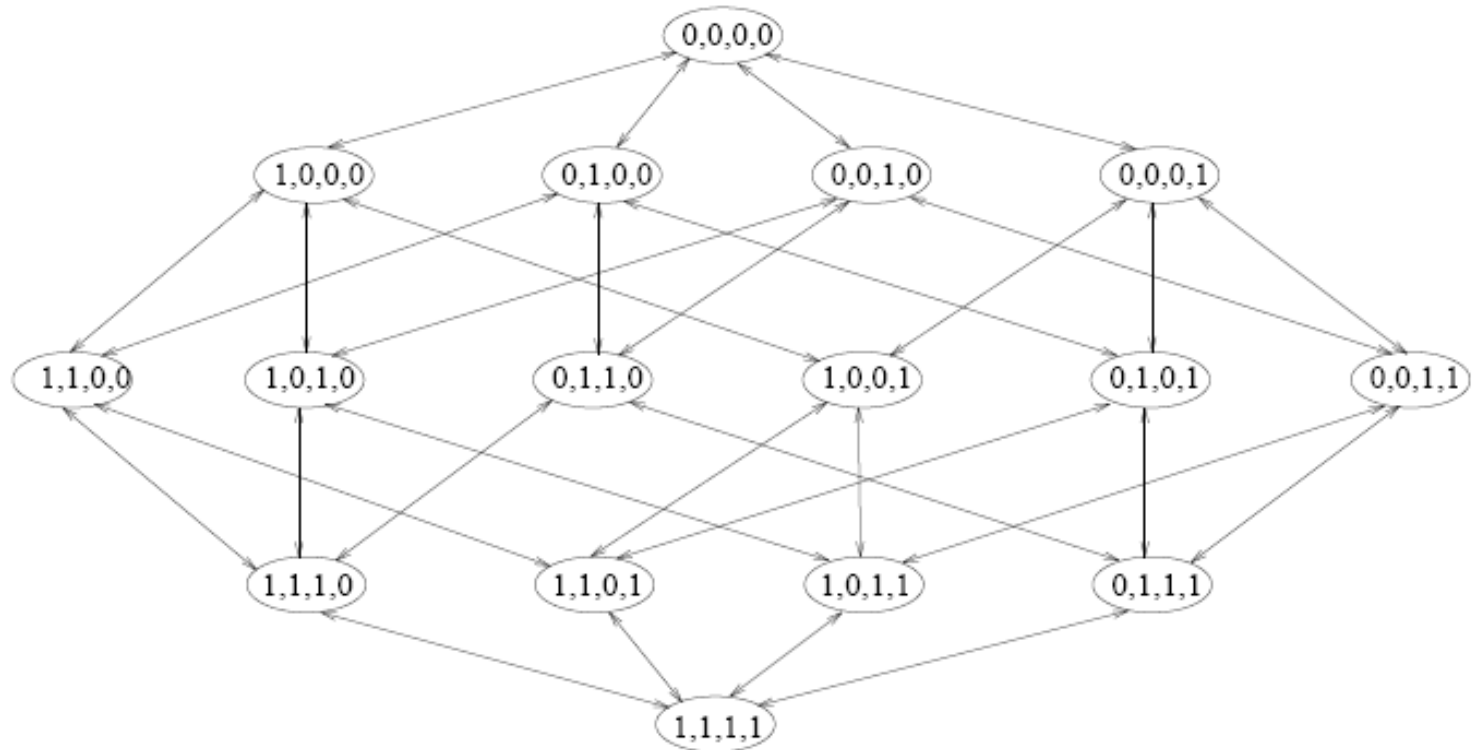


Estrategias wrapper

- Evalúan subconjuntos de objetos utilizando un clasificador.
- Para evitar la búsqueda exhaustiva siguen alguna estrategia de búsqueda.
 - Comúnmente estrategias ávidas o aleatorias

Estrategias wrapper

- Para n objetos, el espacio de búsqueda es de tamaño 2^n





Estrategias wrapper

- Búsqueda exhaustiva
- Búsqueda secuencial
 - Hacia atrás (backward)
 - Hacia adelante (forward)
 - Flotante (floating)
- Búsqueda aleatoria
 - Algoritmos genéticos
 - Búsqueda tabú
 -

Búsqueda exhaustiva

- El tamaño del espacio de búsqueda es 2^n
- Si se busca un número predefinido de objetos el espacio de búsqueda es de tamaño

$$\binom{n}{k}$$

- Para seleccionar 200 objetos de 500 el espacio de búsqueda es de tamaño

$$\binom{500}{200} \approx 5 \times 10^{144} \approx 2^{480}$$

Sequential Forward Selection

Sea D el conjunto de todos los objetos

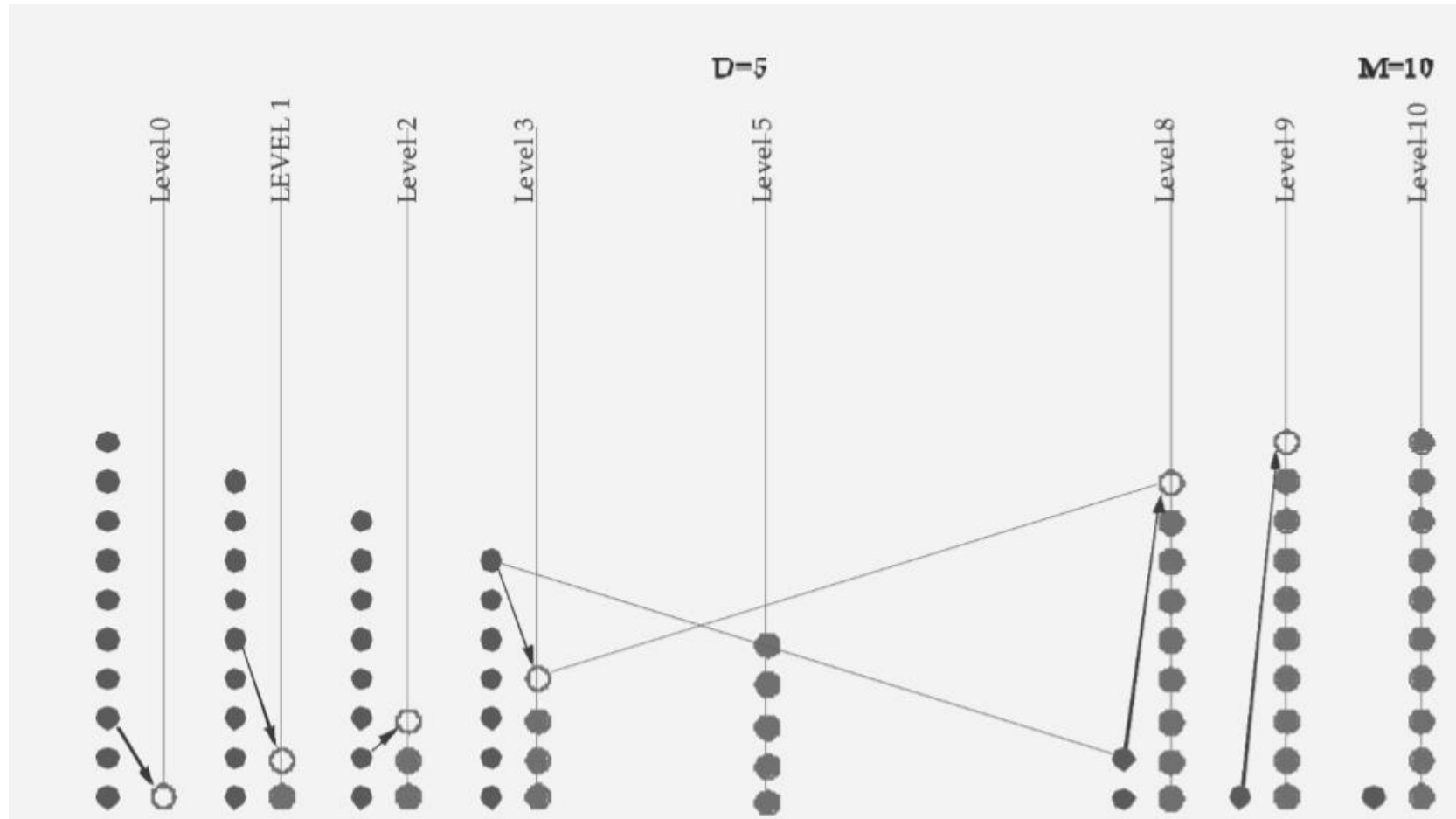
- $S = \emptyset$
- Repetir

$$O = \max_{O \in D \setminus S} (q(S \cup \{O\}))$$

$$S = S \cup \{O\}$$

- Hasta $|S|=k$ / $q(S) < q(S \setminus \{O\})$ / $q(S) > t$

Sequential Forward Selection



Sequential Backward Selection

Sea D en conjunto de todos los objetos

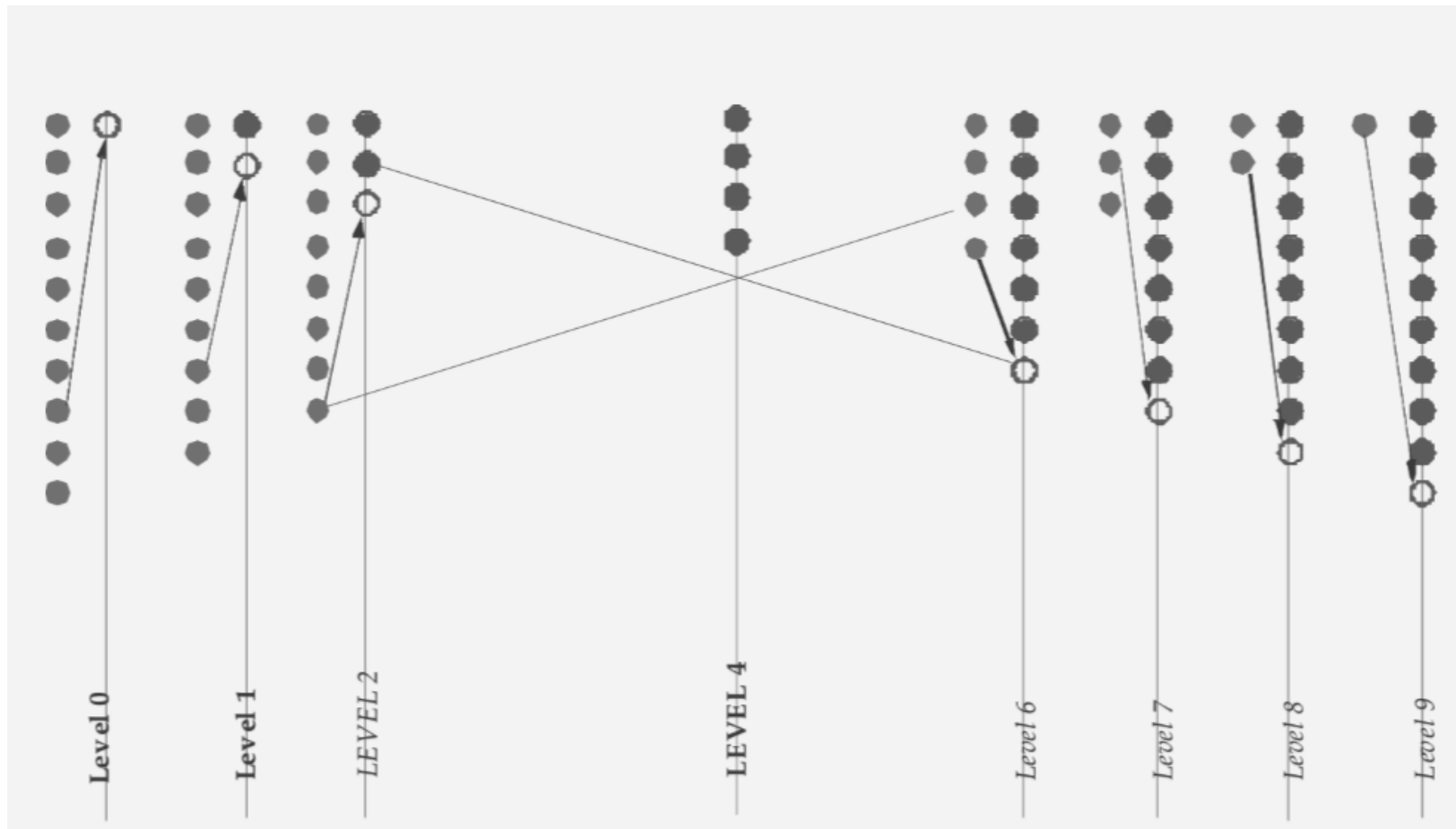
- $S=D$
- Repetir

$$O = \max_{O \in S} (q(S \setminus \{O\}))$$

$$S = S \setminus \{O\}$$

- Hasta $|S|=k$ / $q(S) < q(S \cup \{O\})$ / $q(S) > t$

Sequential Backward Selection



Sequential Floating Forward Selection

- $S = \emptyset$

- Repetir $O = \max_{O \in D \setminus S} (q(S \cup \{O\}))$

$$S = S \cup \{O\}$$

Repetir $O = \max_{O \in S} (q(S \setminus \{O\}))$

$$S = S \setminus \{O\}$$

Hasta $q(S) < q(S \cup \{O\})$

$$S \cup \{O\}$$

- Hasta $|S| = k / q(S) > t$

Sequential Floating Backward Selection

- $S = \emptyset$

- Repetir $O = \max_{O \in S} (q(S \setminus \{O\}))$

$$S = S \setminus \{O\}$$

Repetir $O = \max_{O \in D \setminus S} (q(S \cup \{O\}))$

$$S = S \cup \{O\}$$

Hasta $q(S) \leq q(S \setminus \{O\})$

$S \setminus \{O\}$

- Hasta $|S| = k / q(S) > t$

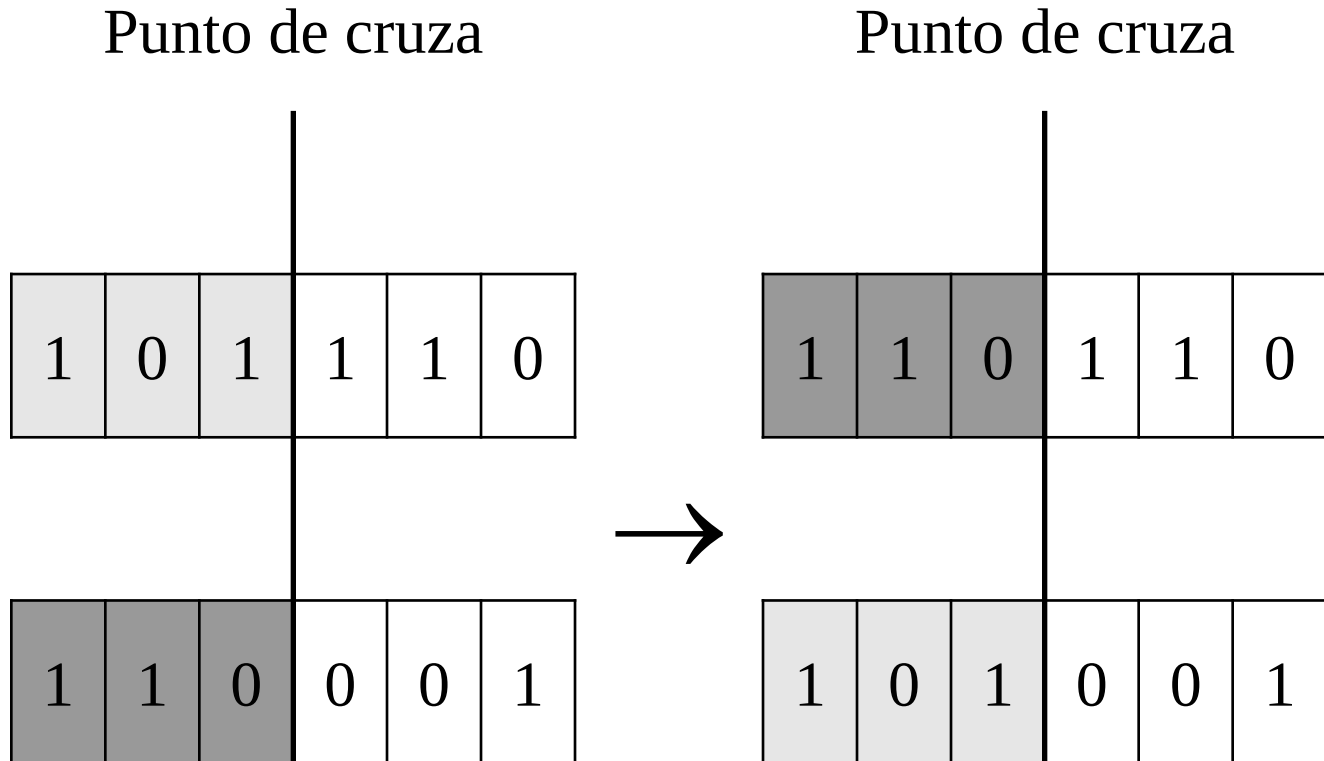
Selección usando Algoritmos Genéticos

- Individuos (suponiendo n objetos)

x_1	x_2	...	x_n
0	1	0/1 ... 0/1	1

Selección usando Algoritmos Genéticos

- Cruza (combinación de individuos)

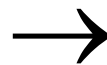


Selección usando Algoritmos Genéticos

- Mutación (alteración de individuos)

Elemento a
mutar

1	1	0	0	0	1
---	---	---	---	---	---



Elemento
mutado

1	0	1	0	0	1
---	---	---	---	---	---

Selección usando Algoritmos Genéticos

- Generar población inicial P
- Para $i=1, \dots, \text{numGeneraciones}$
 - evalúa(P)
 - $P2 = \text{cruza}(P)$
 - evalúa($P2$)
 - $P3 = \text{mutación}(P \cup P2)$
 - evalúa($P3$)
 - $P = \text{selecciona}(P \cup P2 \cup P3)$
- salida = mejorElemento(P)



Evaluación de selectores de objetos

- Utilizando un clasificador
 - Seleccionar un conjunto de bases de datos
 - Utilizar algún método de validación aplicando selección+clasificación
 - Utilizar alguna medida de evaluación de calidad de clasificación



Selección de Objetos

Dr. Jesús Ariel Carrasco Ochoa
ariel@inaoep.mx
Oficina 8311