Tema 4: Técnicas de agrupamiento

Minería de Datos

viu Universidad Internacional de Valencia

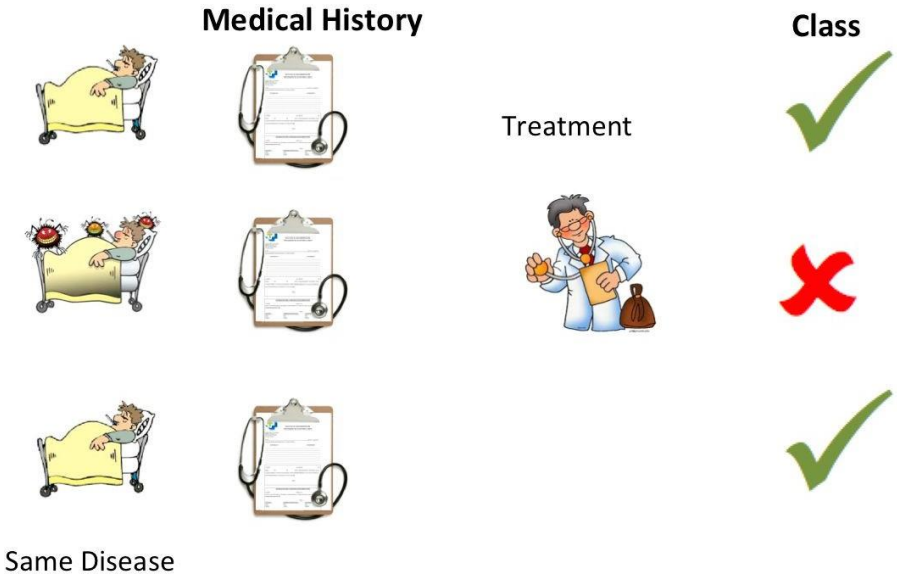
De:
Planeta Formación y Universidades

Índice

- El aprendizaje automático es un tipo de inteligencia artificial (IA) que proporciona a los ordenadores la capacidad de aprender sin ser programados explícitamente. El aprendizaje automático se centra en el desarrollo de programas informáticos capaces de aprender por sí mismos a crecer y cambiar cuando se exponen a nuevos datos.

- Los algoritmos de aprendizaje automático se describen como "supervisados" o "no supervisados".

- En los **algoritmos supervisados**, las **clases** están predeterminadas.
- Estas clases pueden concebirse como un conjunto finito, al que previamente ha llegado un ser humano.



- La tarea del algoritmo es buscar patrones y construir modelos matemáticos.
  - Decision Tree induction.
  - Naive Bayes
  - k-NN (K-nearest neighbour)
  - …

- A continuación, estos modelos se evalúan en función de su capacidad predictiva en relación con las medidas de varianza de los propios datos.

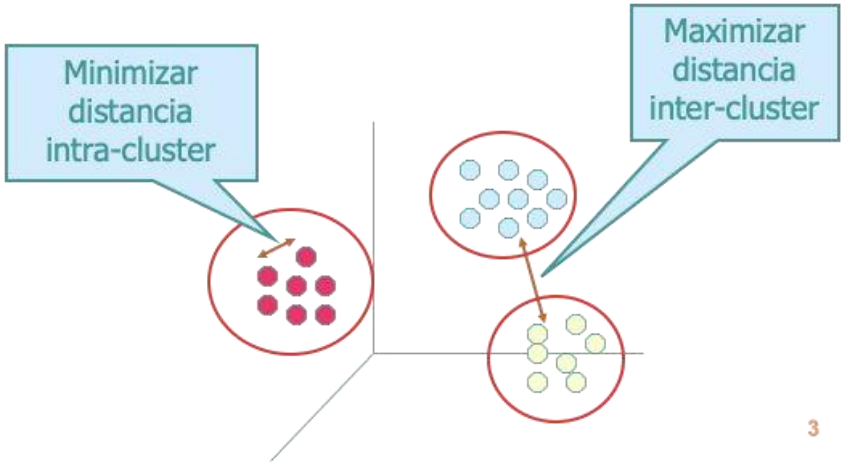
- Un algoritmo de aprendizaje no supervisado **no dispone de clasificadores ni clases.**

- La tarea básica del **aprendizaje no supervisado** es desarrollar etiquetas de clasificación automatizadas.

- Un ejemplo muy común de aplicación de aprendizaje no supervisado son en bases de datos que contengan datos ómicos.

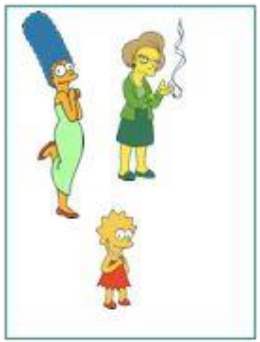
Clustering

Encontrar agrupamientos de tal forma que los objetos de un grupo sean similares entre sí y diferentes de los objetos de otros grupos:



¿Cuál es la forma natural de agrupar los personajes?

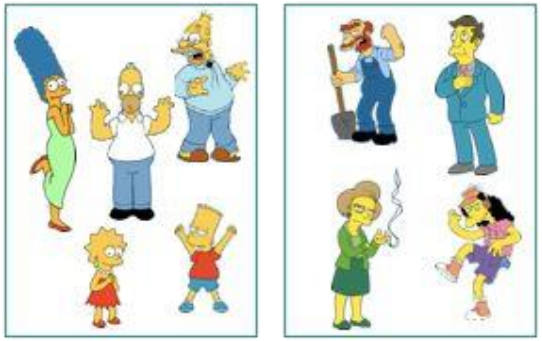
Hombres
vs.
Mujeres



5

¿Cuál es la forma natural de agrupar los personajes?

Simpsons
vs.
Empleados
de la escuela
de Springfield



¿Cuál es la forma natural de agrupar los personajes?

¡¡¡ El clustering es subjetivo !!!

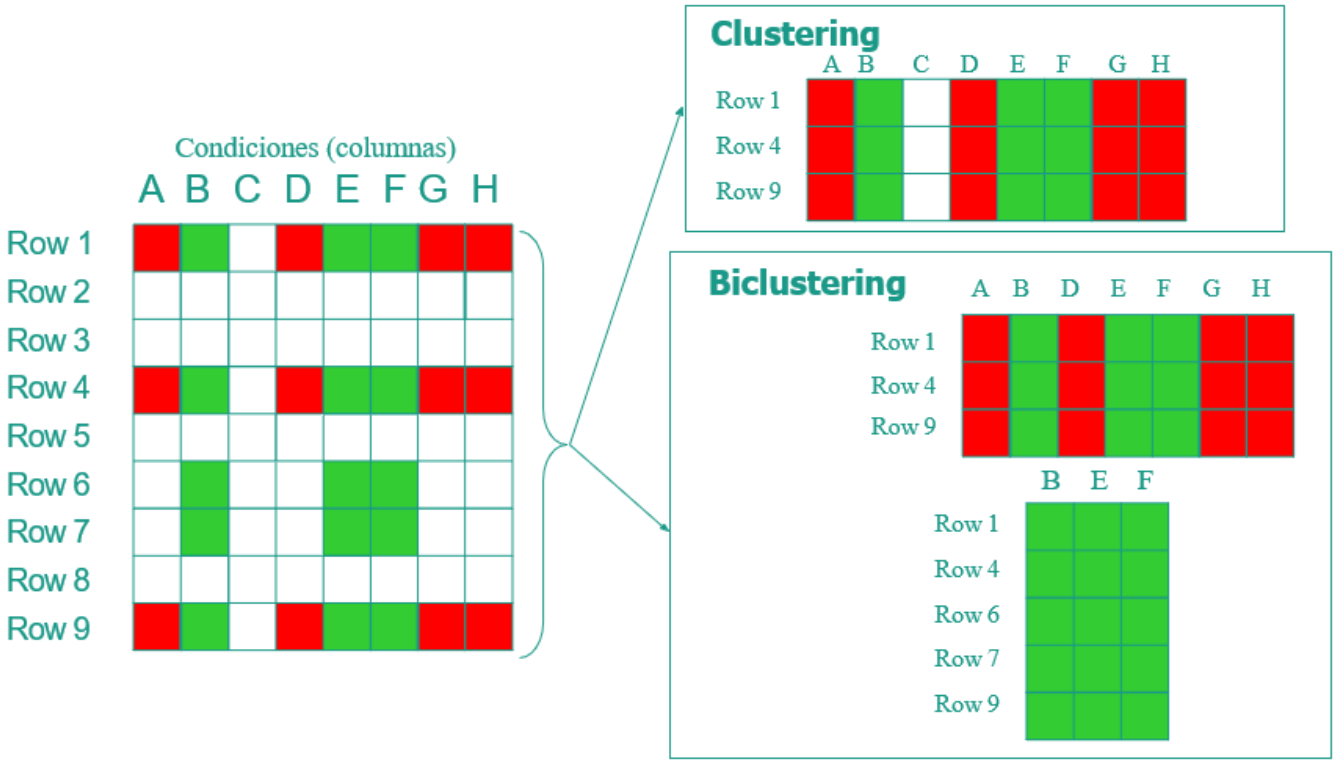
- Primer enfoque descrito por **J.A. Hartigan** en 1972 y usó la terminología 'Direct Clustering'.

- Un ejemplo de aplicación como, por ejemplo, el campo de la biomedicina, se usó el concepto de Clustering por primera vez en el año 2000 por Cheng and Church.

- **Clustering:** Descubre coherencias locales sobre la totalidad de las condiciones (columnas) del conjunto de datos.

- **Biclustering:** Descubre coherencias locales sobre un conjunto de condiciones (columnas) del conjunto de datos.

Aprendizaje no supervisado: Resumen



¿Cómo se parecen unos elementos a otros?

¿Cómo medir la disimilitud de dos elementos?

- Sean a y b vectores, una **medida de distancia** d(a,b) debe obedecer las siguientes reglas:
  - Debe ser positiva $d(a,b) \geq 0 \; \forall a, b \in X$
  - Debe ser simétrica $d(a,b) = d(b,a) \; \forall a, b \in X$
  - Desigualdad triangular $d(a,b) \leq d(a,c) + d(c,b) \; \forall a, b, c \in X$
- Por ejemplo, la disimilitud entre los valores de dos filas: d(fila1,fila2).

Distancias basadas en Minkowski

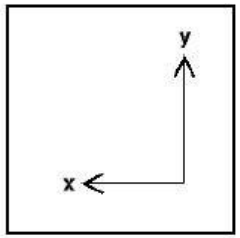
$$d_{Mikowski}(x,y) = \left(\sum_{i=1}^{N} (x_i - y_i)^p \right)^{1/p}$$

- Si **p=1** se le conoce como **medida de distancia Manhattan**:

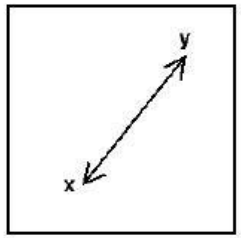
$$d(x,y) = \sum_{i=1}^{m} |x_i - y_i|$$

- Si **p=2**, se le conoce como **distancia euclídea**:

$$d(x,y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$



Manhattan



Euclidean

¡La distancia euclídea no es capaz de distinguir la similitud entre dos valores!

Distancias basadas en Minkowski

- Si **p=infinito** se le conoce como **medida de Chevichef**:

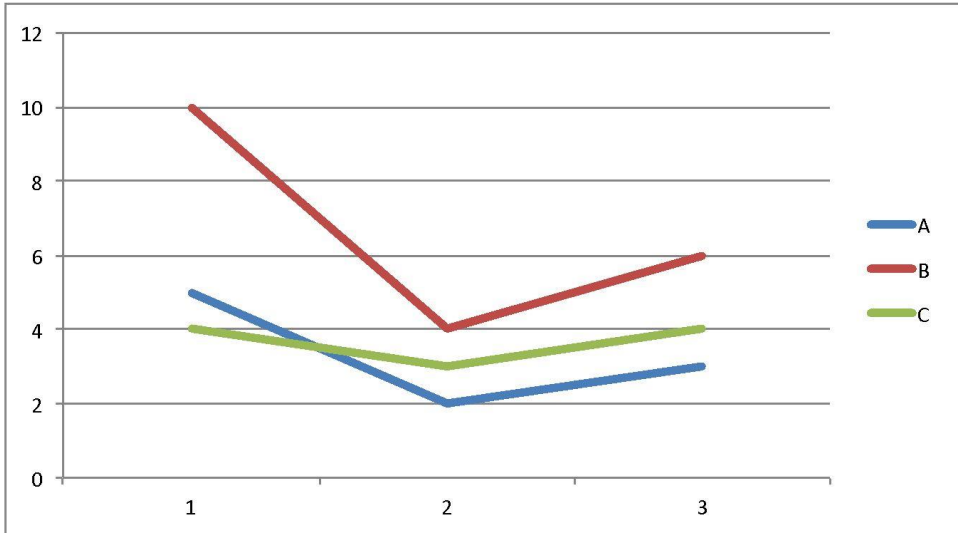
$$d(x,y) = \max_{i=1}^{m} |x_i - y_i|$$

- Esta última medida es popularmente conocida como **distancia del tablero**, ya que en el juego del ajedrez es el número mínimo de movimientos que necesita un rey para ir de una casilla a otra del tablero.

Problema con las medidas de distancia:

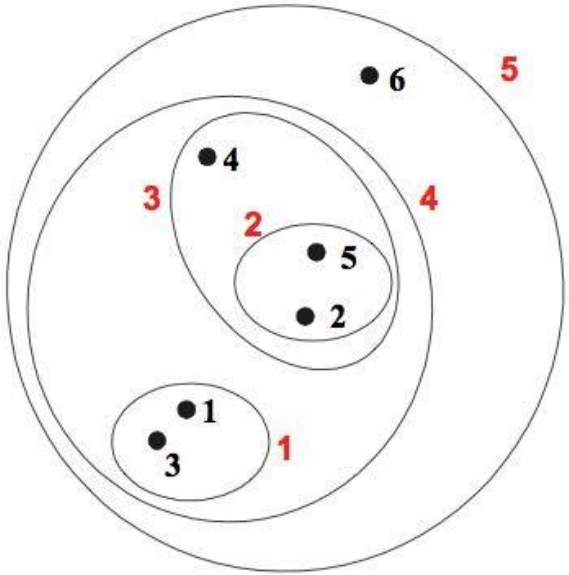
- Analizando los patrones de comportamiento, el problema radica cuando se deben enfrentar a un conjunto de valores (vectores):

A	B	C
5	10	4
2	4	3
3	6	4

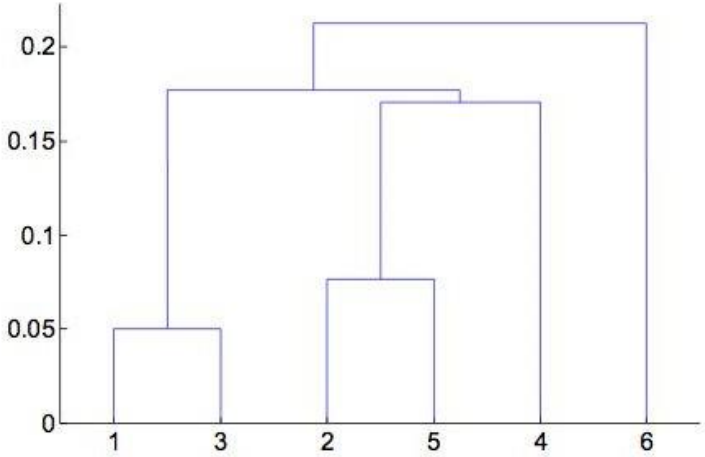


Medidas de similitud

- Las medidas de similitud se usan cuando se pretende medir lo parecidos que son **dos vectores**. Estas medidas suelen devolver un valor que se encuentra entre [-1,1] y su significado es el siguiente:
  - valorMedida=1, significa que hay una correlación directa.
  - valorMedida=0, significa que no hay correlación.
  - valorMedida=-1, significa que existe una correlación inversa.
- Las medidas de similitud más conocidas son: **Pearson** y **Spearman** entre otros.
- **Pearson:** Los datos deben seguir una distribución normal y hay que tener especial cuidado con los valores atípicos. Sólo miden relaciones lineales.
- **Spearman:** Sólo miden relaciones monótonas.

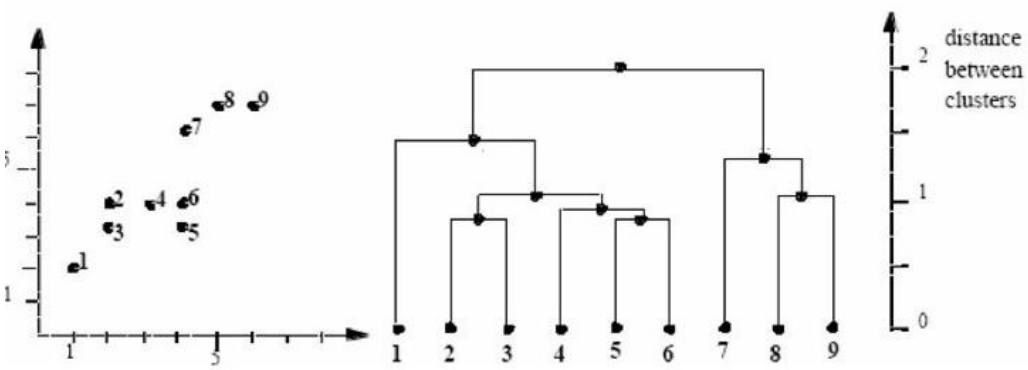


Hierarchical Clustering



Dendrogram

Dendograma



- La raíz *(root)* representa todo el conjunto de datos.
- Una hoja *(leaf)* representa un único objeto del conjunto de datos.
- Un nodo interno *(internal node)* representa la unión de todos los objetos de su subárbol.
- El peso *(weight)* de un nodo interno representa la distancia entre sus dos nodos hijos.

Enfoques

Aglomerativo:

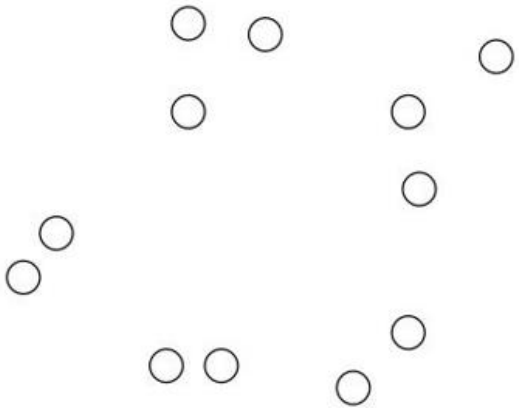
- Inicio: Cada punto es un cluster individual.
- Por cada paso: Se fusiona un par de clusters más cercanos.
- Finaliza: Hasta que sólo quede un cluster o k clusters.

Divisivo:

- Inicio: Todos los puntos forman un único cluster
- Por cada paso: Se divide un cluster.
- Finaliza: Hasta que un cluster contenga un punto (o haya k clusters).
- Hay que decidir qué cluster dividir en cada paso.

Situación de inicio

Para el clustering jerárquico aglomerativo partimos de clusters de puntos individuales y de una matriz de proximidad.

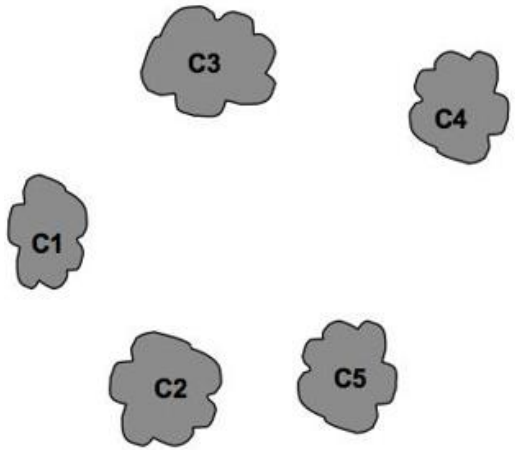


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						

Matriz de proximidad

Situación intermedia

Después de algunos pasos de fusión, tenemos algunos clusters.

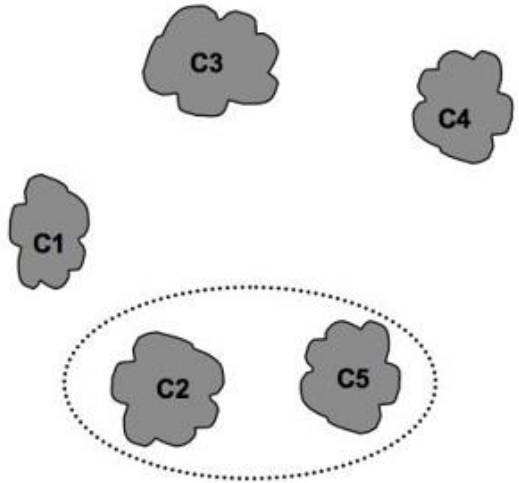


	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Matriz de proximidad

Situación intermedia

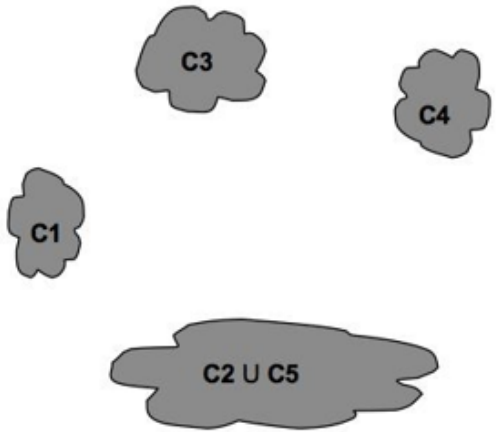
Queremos unir los dos clusters más cercanos (C2 y C5) y actualizar la matriz de proximidad.



	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Matriz de proximidad

Después de la unión: ¿cómo actualizamos la medida de proximidad?



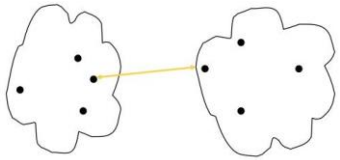
	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Matriz de distancia

- La operación clave es el **cálculo de la distancia** entre dos clusters.
- Los distintos algoritmos se distinguen por diferentes enfoques a la hora de definir la distancia entre clusters.

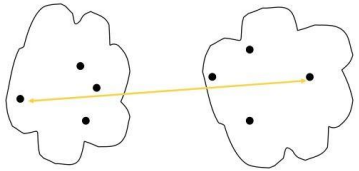
Ejemplos de cálculo de la distancia entre dos clusters

Único



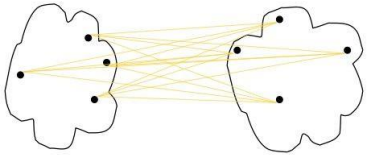
$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} \{ d(x, y) \}$$

Completo



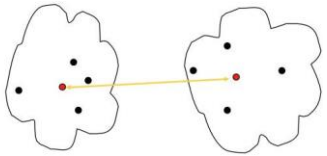
$$d(C_i, C_j) = \max_{x \in C_i, y \in C_j} \{ d(x, y) \}$$

Medio



$$d(C_i, C_j) = \frac{1}{|C_i| |C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$$

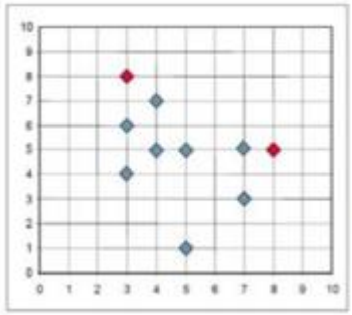
Centroide



$$d(C_i, C_j) = d(c_i, c_j)$$

$$c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \qquad c_j = \frac{1}{|C_j|} \sum_{x \in C_j} x$$

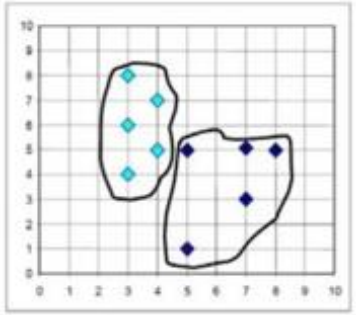
K-means



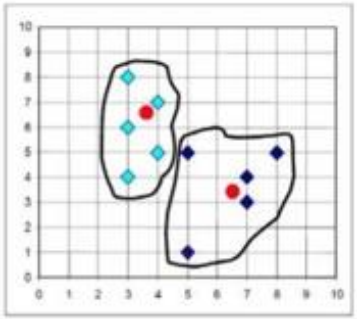
K=2

Elegir arbitrariamente K puntos como centroides iniciales

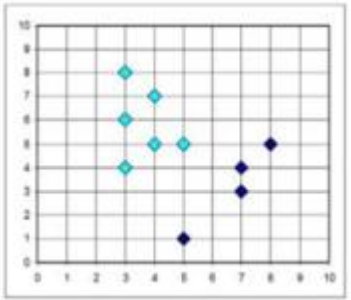
Asignar puntos al centroide más cercano



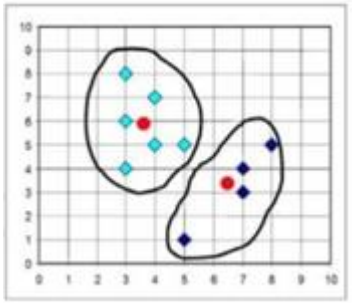
Actualizar el centroide del clúster



Asignar puntos

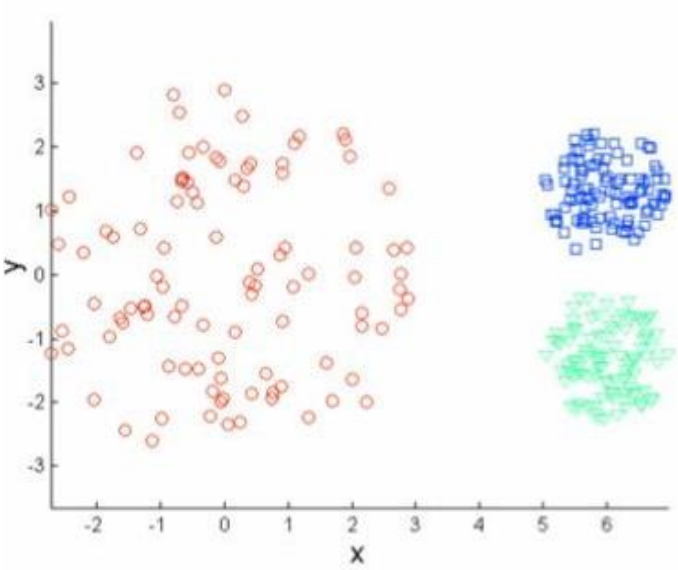


Actualizar el centroide del clúster

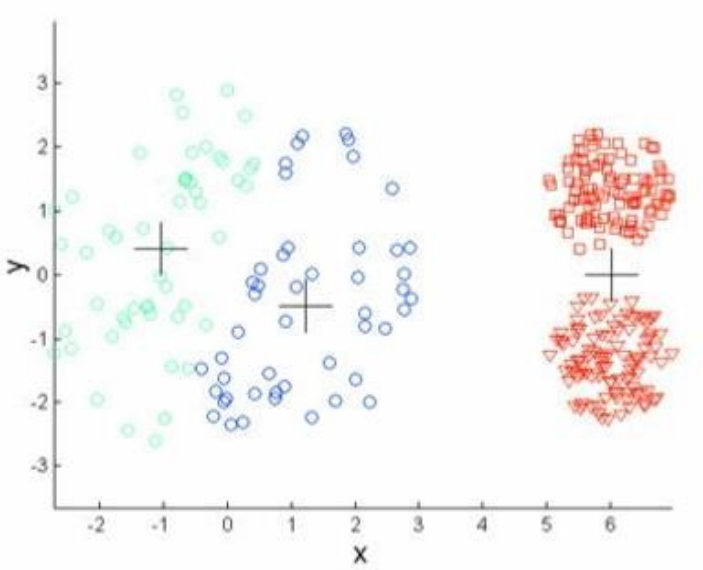


Asignar puntos

Limitación: Densidad diferenciadas

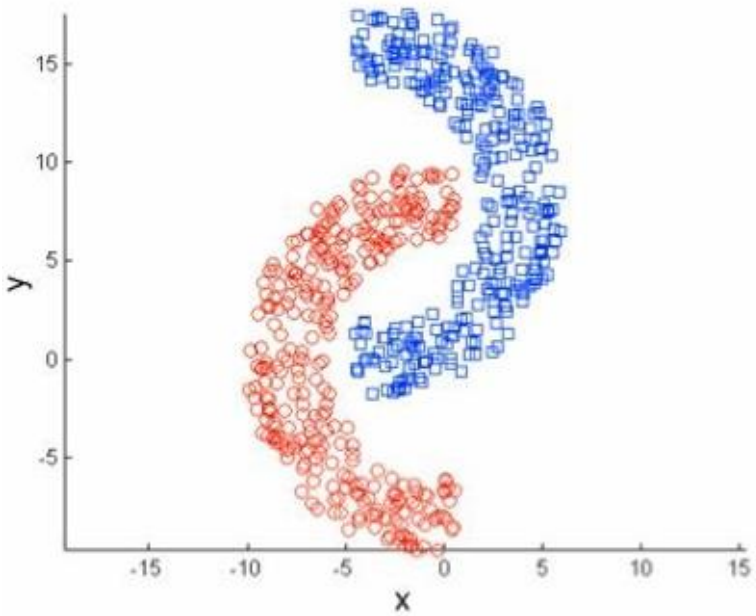


Puntos originales

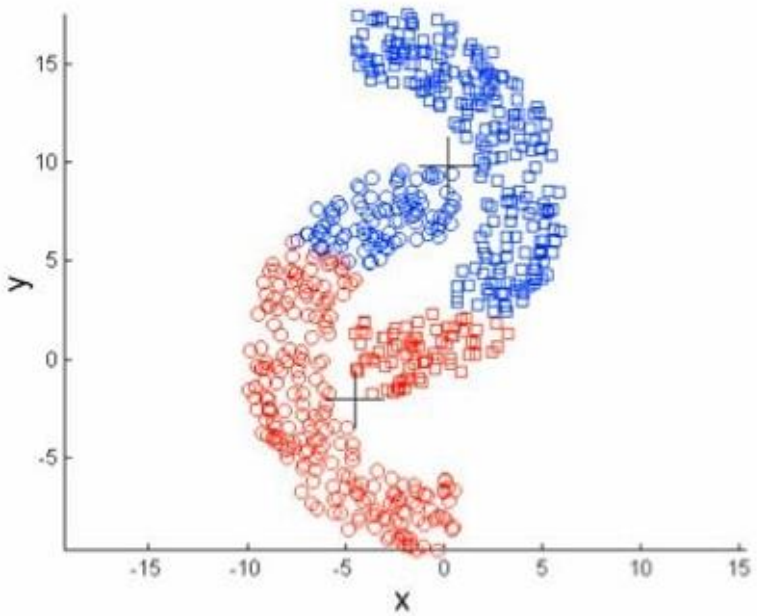


K-Means (3 clusters)

Limitación: Formas no globulares

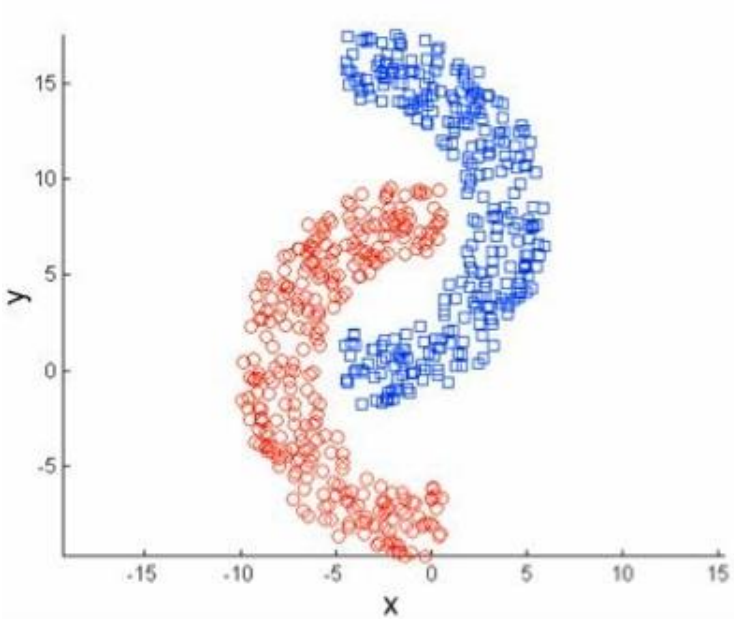


Puntos originales

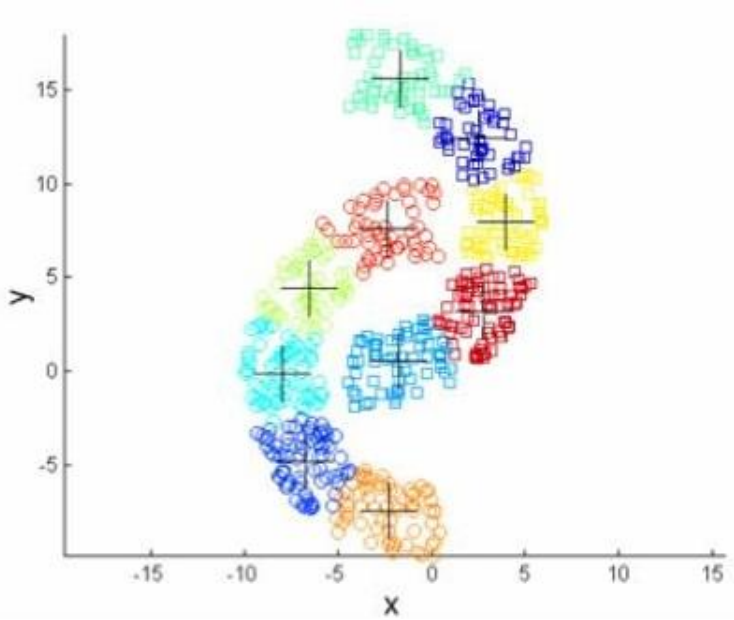


K-Means (2 clusters)

Limitación: Superación



Puntos originales



K-Means clusters

¿Por qué Biclustering?

- En Clustering existe una clara dependencia ya que se generan clusters de filas en el que sus elementos tienen el mismo comportamiento para todas las columnas del dataset, o viceversa.

- En muchas aplicaciones no se desea que se construyan clusters en el que se base en todas las condiciones (columnas) o todas las filas del dataset.

- Ejemplos:
  - En genómica, no todos los genes se comportan de la misma manera para todos los pacientes con una determinada enfermedad.
  - En eficiencia energética, cuando se desea obtener un patrón de comportamiento común de una serie temporal en el que influyen diferentes máquinas de consumo energético.
  - etc.

¿Inconvenientes de Clustering en Big Data?

- A medida que aumenta el número de columnas, es cada vez menos probable que ciertas filas conserven la similitud en todas las columnas. Por lo que la agrupación puede resultar difícil.

- Clustering no ofrece todo el conocimiento oculto de los datos, ya que descartan aquellas relaciones presentes que se encuentren en algunos atributos (columnas) y que pueden llegar a ser realmente significativas.

Tipos de biclusters

1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0

Bicluster constante

1.0	1.0	1.0	0.0
2.0	2.0	2.0	2.0
3.0	3.0	3.0	3.0
4.0	4.0	40	40

Con filas constantes

1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0

Con columnas constantes

1.0	2.0	5.0	0.0
2.0	3.0	6.0	1.0
4.0	5.0	8.0	3.0
5.0	6.0	9.0	4.0

Valores coherentes. Modelo aditivo.

1.0	2.0	0.5	1.5
2.0	4.0	1.0	3.0
4.0	8.0	2.0	6.0
3.0	6.0	1.5	4.5

Valores coherentes. Modelo multiplicativo

S1	S1	S1	S1
S1	S1	S1	S1
S1	S1	S1	S1
S1	S1	S1	S1

Bicluster de evolución coherente general

S1	S1	S1	S1
S2	S2	S2	S2
S3	S3	S3	S3
S4	S4	S4	S4

Evolución coherente en filas

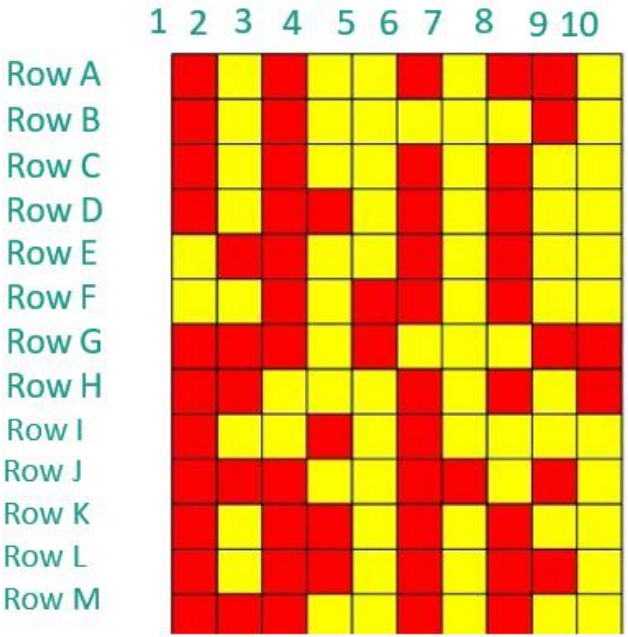
S1	S2	S3	S4
S1	S2	S3	S4
S1	S2	S3	S4
S1	S2	S3	S4

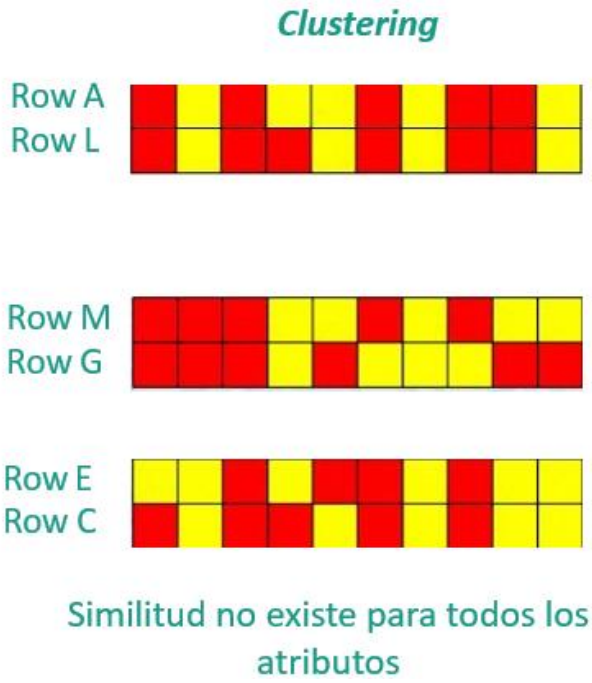
Evolución coherente en columnas

70	13	19	10
29	40	49	35
40	20	27	15
90	15	20	12

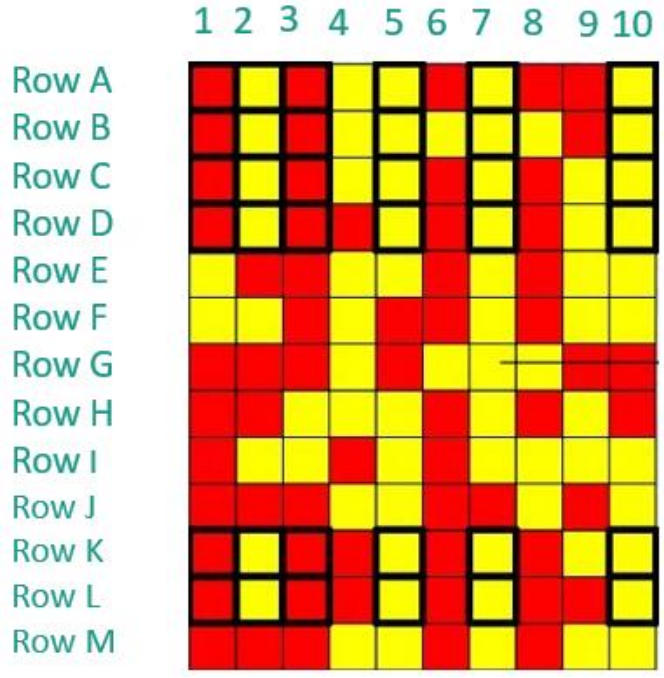
Evolución coherente en columnas

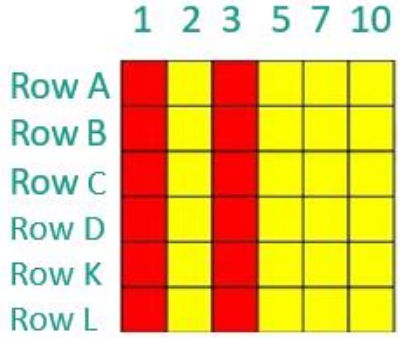
Biclustering vs Clustering





Biclustering vs Clustering





Bicluster {1,2,3,5,7,10} {A,B,C,D,K,L}

Enfoques de Biclustering

Algoritmo BiBit

Aurelio López-Fernández, Domingo Rodríguez-Baena, Francisco A. Gómez Vela, Federico Divina, Miguel García (2021). A multi-GPU biclustering algorithm for binary datasets. *Journal of Parallel and Distributed Computing*, *147*, 209-219.

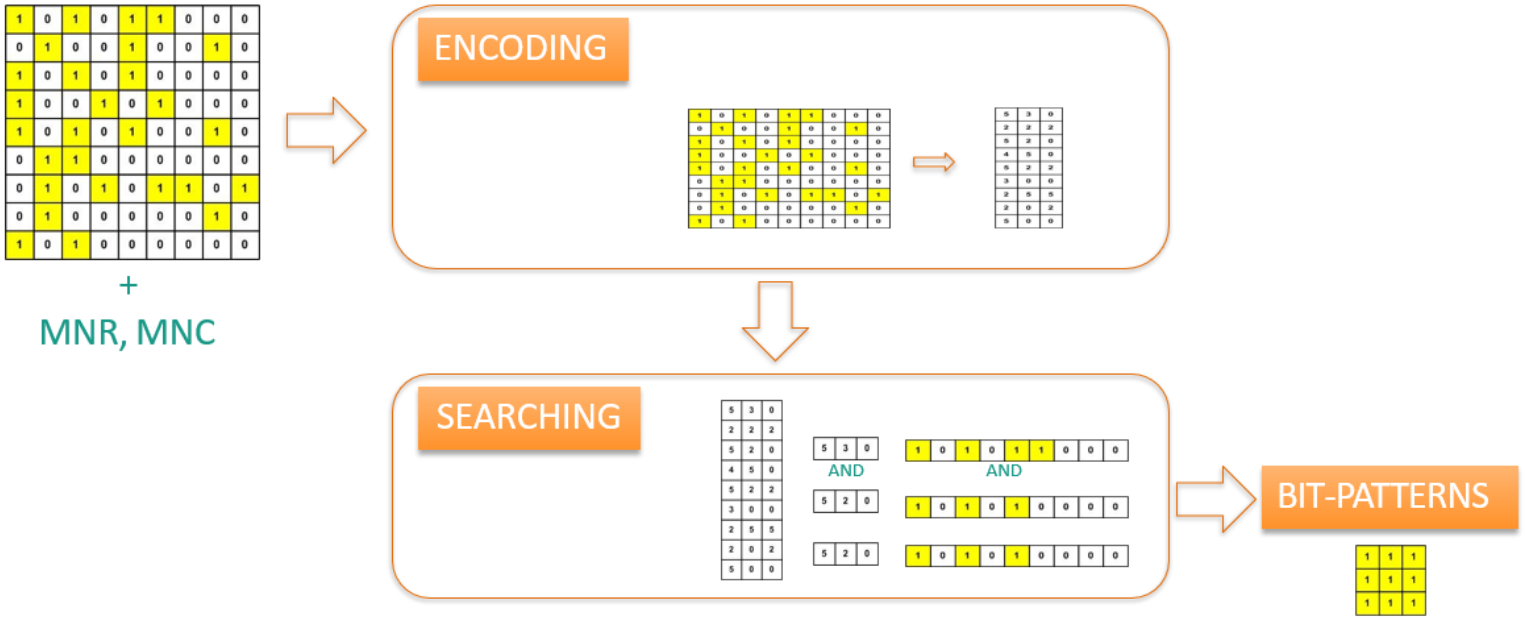
Algoritmo Evo-Bexpa

Beatriz Pontes, Raúl Giráldez, Jesús S. Aguilar-Ruiz (2013). Configurable pattern-based evolutionary biclustering of gene expression data. *Algorithms for Molecular Biology*, *8*(1), 1-22.

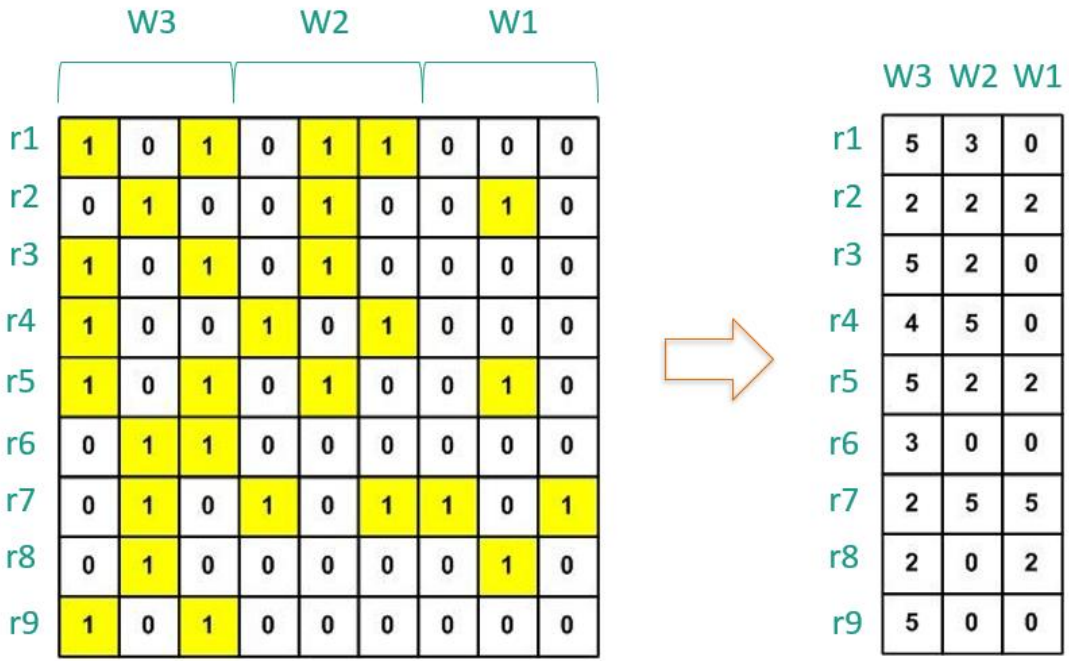
Algoritmo BBCF

Sun, J., & Zhang, Y. (2022). Recommendation System with Biclustering. *Big Data Mining and Analytics*, *5*(4), 282-293.

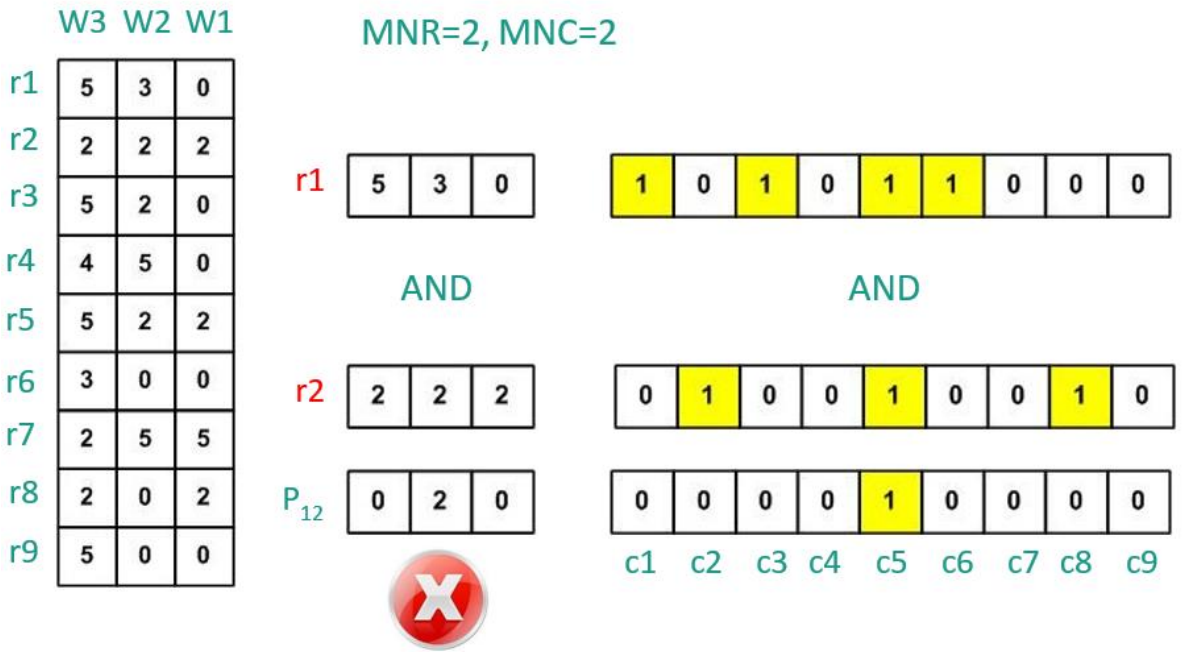
Metodología



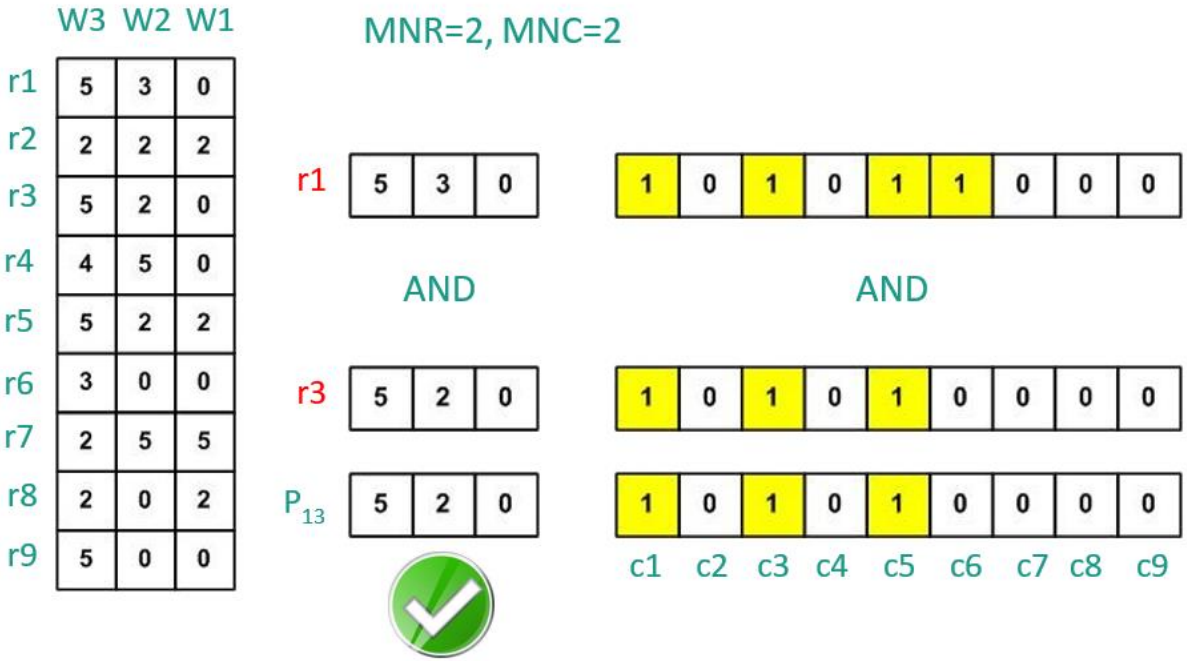
Metodología: Fase de codificación



Metodología: Fase de búsqueda



Metodología: Fase de búsqueda



Metodología: Fase de búsqueda

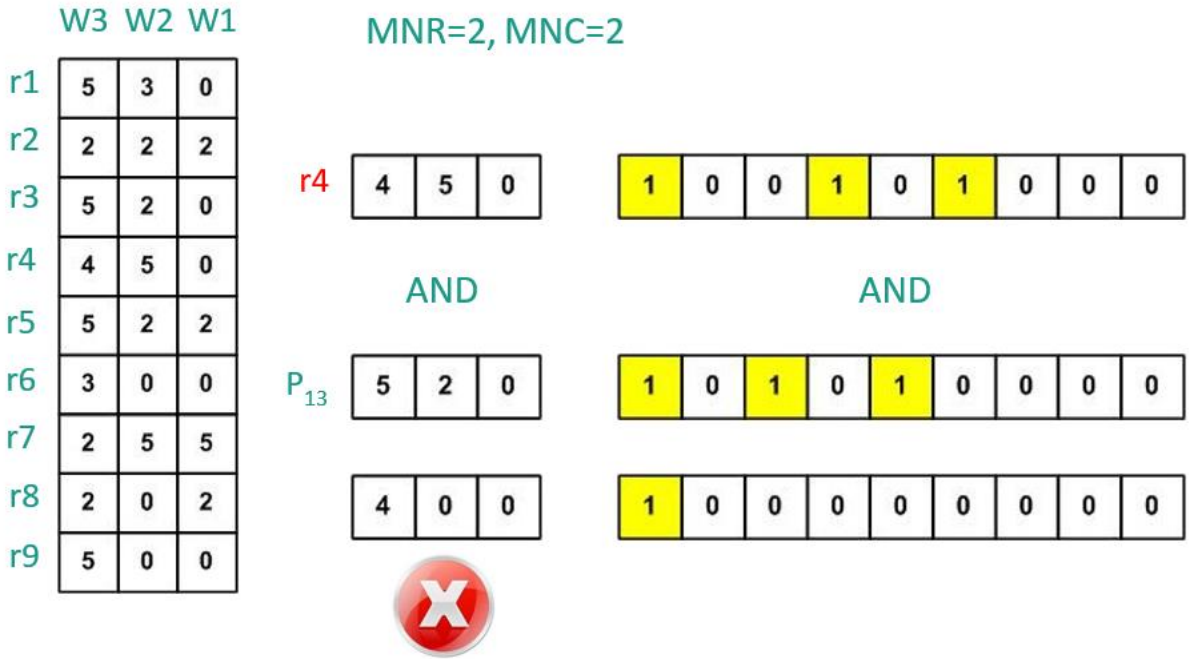
	W3	W2	W1
r1	5	3	0
r2	2	2	2
r3	5	2	0
r4	4	5	0
r5	5	2	2
r6	3	0	0
r7	2	5	5
r8	2	0	2
r9	5	0	0

MNR=2, MNC=2

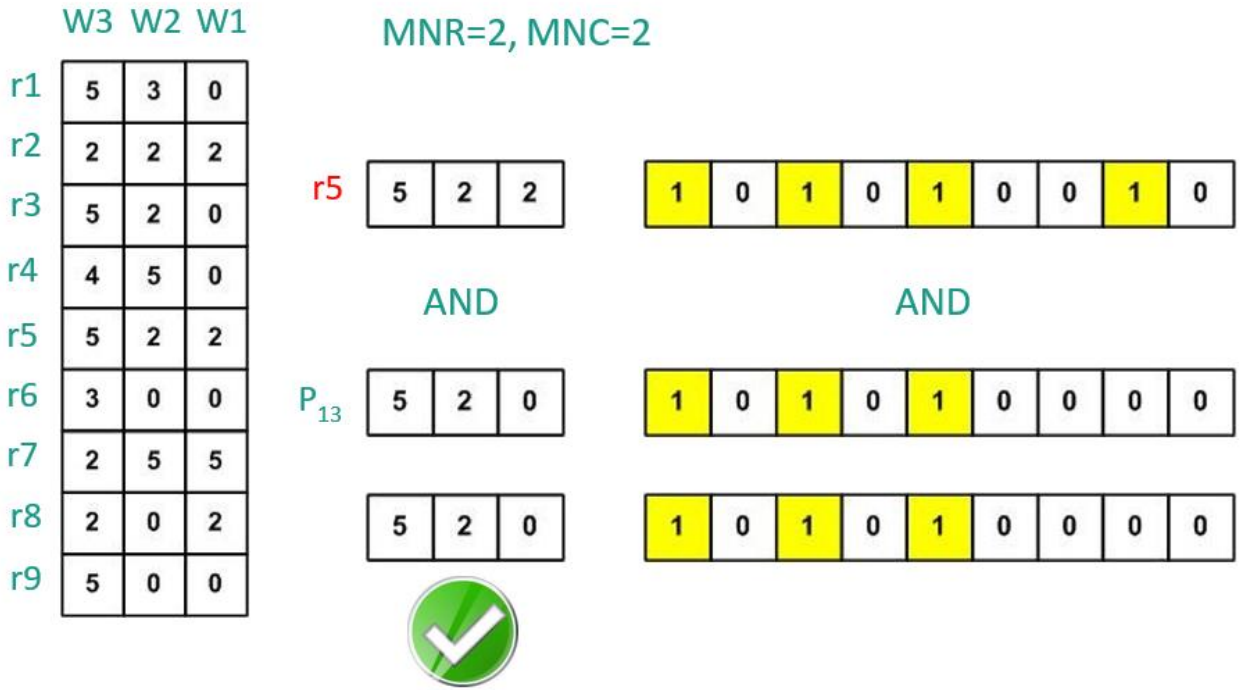
B_{13}={{r1,r3}, {c1,c3,c5}}

	c1	c3	c5
r1	1	1	1
r3	1	1	1

Metodología: Fase de búsqueda



Metodología: Fase de búsqueda



Metodología: Fase de búsqueda

	W3	W2	W1
r1	5	3	0
r2	2	2	2
r3	5	2	0
r4	4	5	0
r5	5	2	2
r6	3	0	0
r7	2	5	5
r8	2	0	2
r9	5	0	0

MNR=2, MNC=2

B_{13}={{r1,r3,r5}, {c1,c3,c5}}

	c1	c3	c5
r1	1	1	1
r3	1	1	1
r5	1	1	1

Gracias

viu Universidad Internacional de Valencia

De:
Planeta Formación y Universidades