Clasificación NO SUPERVISADA

AGRUPAMIENTO

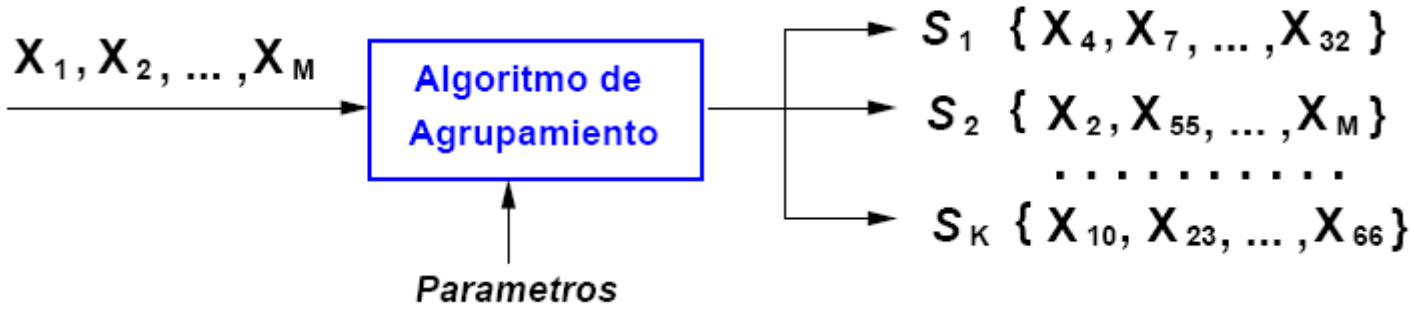
Clasificación No Supervisada

Se trata de construir clasificadores sin información a priori, o sea, a partir de conjuntos de patrones no etiquetados

Objetivo: Descubrir la estructura de los datos, buscar agrupamientos

Agrupamiento

- Técnica diseñada para realizar una clasificación asignando patrones a grupos (clusters) de tal forma que cada grupo sea más o menos homogeneo y distinto de los demás
- Criterio de homogeneidad: distancia

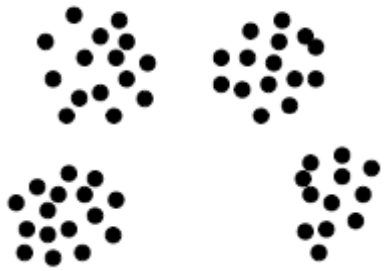


Agrupamiento

- ¿De qué depende el resultado de un algoritmo de agrupamiento?
  - Del algoritmo concreto empleado
  - El valor de los parámetros del algoritmo
  - Los patrones utilizados y el orden en que se procesan
  - La medida de similaridad adoptada

Agrupamiento

Ejemplo:

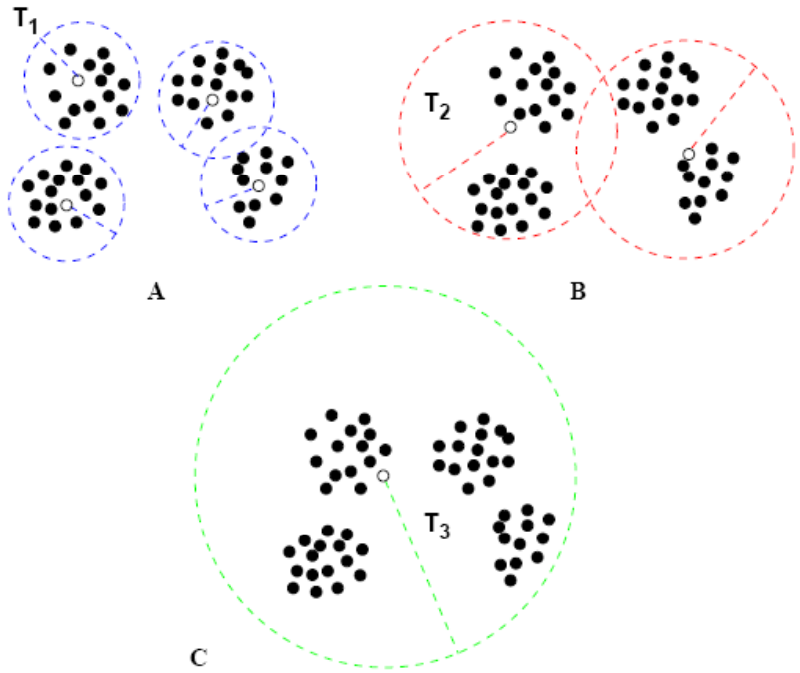


Patrones a agrupar

Algoritmo:

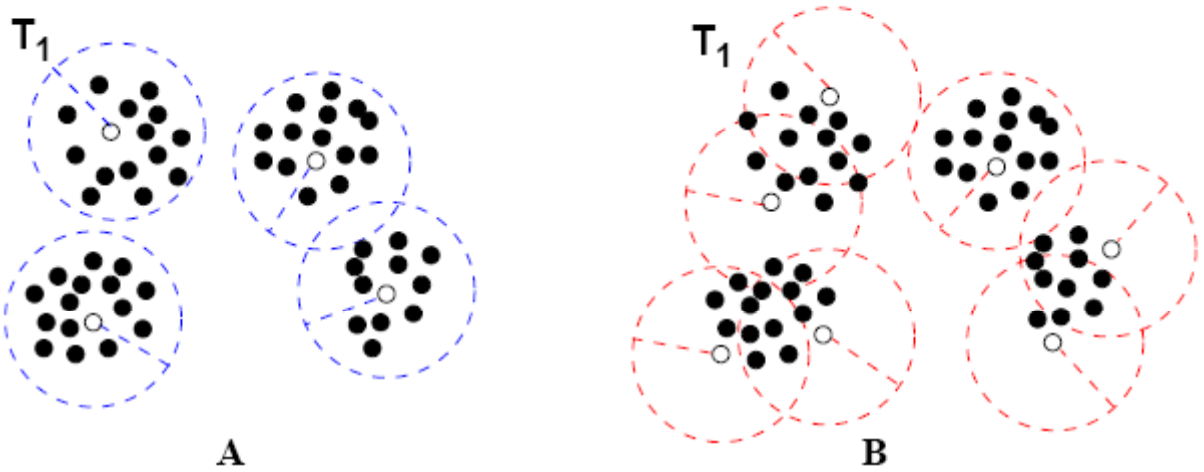
1. Selección de un patrón como el centro de un agrupamiento
2. Se procesan los restantes patrones de manera que se crea un nuevo agrupamiento si la distancia del patrón al agrupamiento más cercano es mayor que T o bien lo añade a un agrupamiento existente si es menor que T

Agrupamiento



Dependencia del umbral T

Agrupamientos



Dependencia del orden

Medidas de Similaridad

- Distancia de Mahalanobis:

$$\delta_M^2(X_i, X_j) = (X_i - X_j)^T \sum^{-1} (X_i - X_j)$$

La covarianza considera la distinta dispersión de las variables en el espacio

- Distancia Euclídea:

$$\delta_E^2(X_i, X_j) = (X_i - X_j)^T (X_i - X_j) = \sum_{k=1}^{d} (X_{ik} - X_{jk})^2$$

Medidas de Similaridad

- Distancia Euclídea ponderada:

$$\delta^2_{E\sigma}(X_i, X_j) = \sum_{k=1}^{d} \alpha_k (X_{ik} - X_{jk})^2$$

Siendo $\alpha_k = \frac{1}{\sigma^2_{mk}}$ $\mu_m = \begin{bmatrix} \mu_{m_1} \\ \mu_{m_2} \\ \vdots \\ \mu_{m_d} \end{bmatrix}$ $\sigma^2_m = \begin{bmatrix} {\sigma_{m_1}}^2 \\ {\sigma_{m_2}}^2 \\ \vdots \\ {\sigma_{m_d}}^2 \end{bmatrix}$

Para el agrupamiento Sm

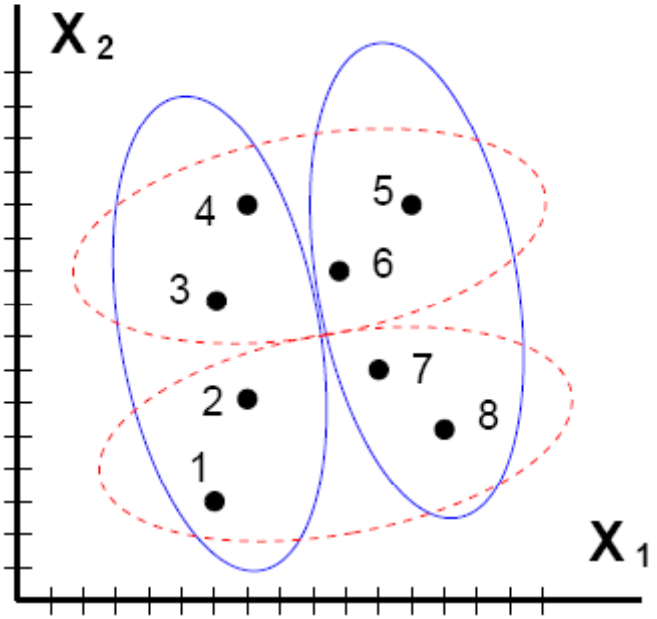
Medida de la calidad del agrupamiento

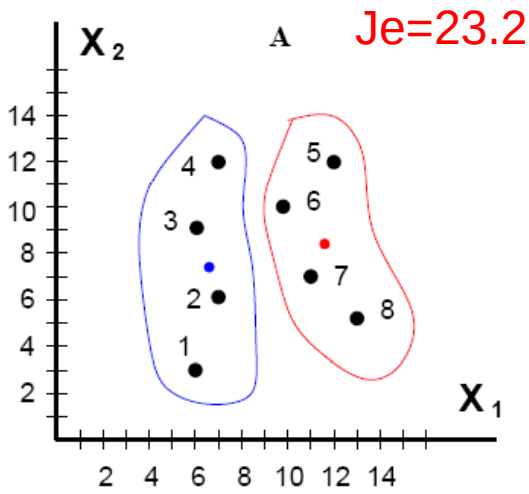
- ¿Cuál es la mejor agrupación de entre todas las posibles?
- Función criterio más empleada: Suma de los errores al cuadrado, son las particiones de mínima varianza

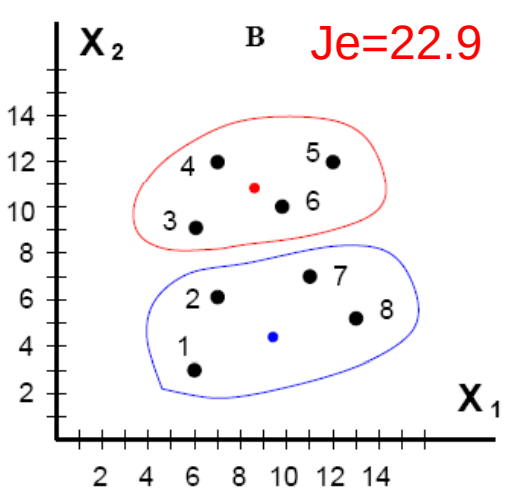
$$Z_i = \frac{1}{N_i} \sum_{X \in S_i} X \qquad J_e = \sum_{i=1}^{K} \sum_{X \in S_i} ||X - Z_i||^2$$

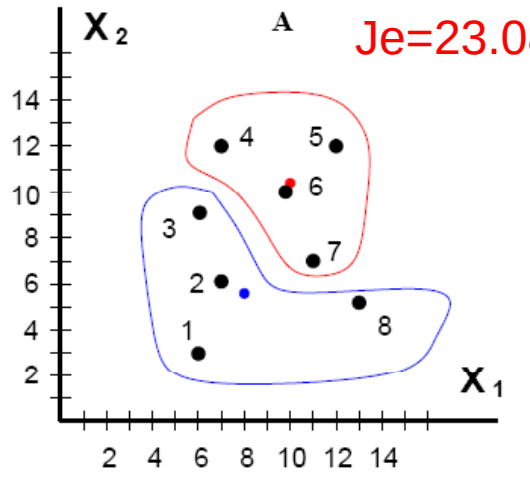
Medida de la calidad del agrupamiento

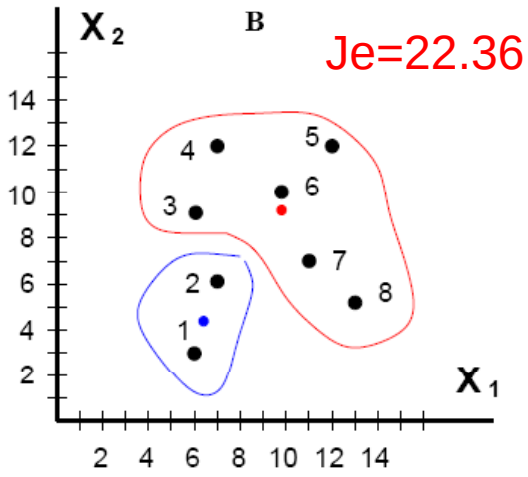
- Ejemplo:











Algoritmos de Agrupamiento

- Agrupamiento: problema de optimización en el que se busca la partición que optimiza la función criterio.
- Búsqueda exhaustiva inabordable

$$K^M / K!$$

Para M patrones con K agrupamientos

Algoritmos de Agrupamiento

- En función de si hay función criterio a optimizar:
  - Algoritmos directos o heurísticos (no optimizan ninguna función criterio)
  - Algoritmos indirectos o por optimización si se usa función criterio

Algoritmos de Agrupamiento

- Según la construcción del agrupamiento:
  - Algoritmos incrementales (‘bottom-up’). Parten de patrones aislados y tienden a unir agrupamientos de acuerdo a algún umbral
  - Algoritmos decrementales (‘top-down’). Parten de agrupamientos ya establecidos y crean nuevos agrupamientos más homogeneos
  - Algoritmos mixtos

Algoritmos de Agrupamiento

- Según si se conocen o no el número de agrupamientos:
  - Número de clases desconocido:
    - Método adaptativo
    - Algoritmo de Batchelor y Wilkins
  - Número de clases conocido:
    - Algoritmo de las K-medias
    - Agrupamiento secuencial
    - Algoritmo ISODATA

Algoritmos de Agrupamiento

- Número de clases desconocida
  - Método Adaptativo
  - Algoritmo de Batchelor y Wilkins
- Número de clases conocida
  - Algoritmo de las K-medias
  - Agrupamiento secuencial
  - Algoritmo ISODATA

Método Adaptativo

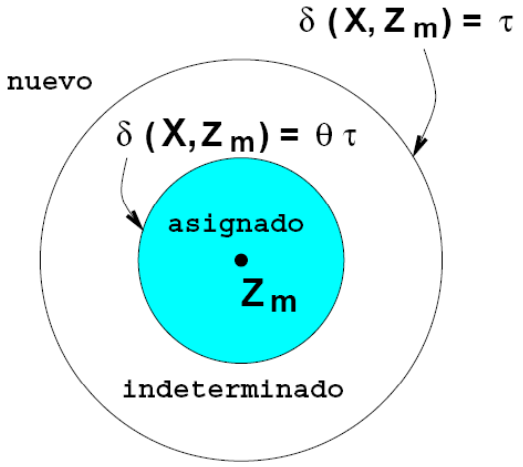
- El número de clases se desconoce
- Método heurístico con dos parámetros básicos:
  - Umbral de distancia para crear agrupamientos τ
  - Fracción de τ que determina total confianza θ

Método Adaptativo

Algoritmo

Se usa una parte M del total de patrones a agrupar

1. Inicialización: Se crea el primer agrupamiento con un conjunto de datos formando X_1
2. Asignación inicial: Se establecen los primeros agrupamientos evaluando para cada X



Si $\delta\ (X, Z_m) > \tau$
    $estado =$ nuevo
Si-no
    Si $\delta\ (X, Z_m) \leq \theta\tau$
        $estado =$ asignado
    Si-no
        $estado =$ indeterminado
    Fin-si
Fin-si

Método Adaptativo

3. Reasignar hasta estabilización: Volver a procesar X1, X2, …, XN. Los patrones pueden cambiar de estado debido a las actualizaciones de los centros de los agrupamientos. Este proceso se repite hasta que no haya cambios de estado
4. Agrupar libremente: Se agrupan los M-N patrones no empleados hasta ahora según la regla de mínima distancia. No se considera la región de indeterminación, pero se incorpora la región de rechazo.

Método Adaptativo

- Ventajas:
  - Método sencillo, sólo dos parámetros
  - Eficiente, escaso cálculo computacional
- Inconvenientes:
  - Produce agrupamientos compactos y separados de los demás
  - El resultado está sesgado por los primeros patrones utilizados en el aprendizaje, depende del orden de los patrones

Algoritmos de Agrupamiento

- Número de clases desconocida
  - Método Adaptativo
  - Algoritmo de Batchelor y Wilkins
- Número de clases conocida
  - Algoritmo de las K-medias
  - Agrupamiento secuencial
  - Algoritmo ISODATA

Algoritmo de Batchelor y Wilkins

- Método heurístico incremental con un único parámetro: θ (fracción de la distancia media entre los agrupamientos existentes).
- Ideas generales:
  - Se establece un agrupamiento si la distancia de un patrón al agrupamiento más cercano supera un umbral. El primer agrupamiento se establece arbitrariamente.
  - El umbral de distancia ahora no es fijo y se calcula en base al parámetro θ y a la distancia media entre los agrupamientos existentes en el momento de su evaluación
  - El aprendizaje termina cuando no se crean nuevos agrupamientos

Algoritmo de Batchelor y Wilkins

- **CalculaUmbral** (θ): devuelve el valor umbral de distancia por encima del cual se creará un nuevo agrupamiento. Este valor se calcula ponderando la distancia media entre los centros de los A agrupamientos existentes:

$$
\begin{aligned}
&sum = 0 \\
&\text{Repite para } i = 1, 2, \ldots, A-1 \\
&\qquad sum = sum + \delta(Z_i, Z_{i+1}) \\
&\text{Fin-para} \\
&umbral = \theta \ \frac{sum}{A-1} \\
&\text{Devuelve } (umbral)
\end{aligned}
$$

Algoritmo de Batchelor y Wilkins

1. Inicialización: Se establecen los centros de los dos primeros agrupamientos. El primero arbitrariamente (primer patrón), el segundo el más alejado del primero
2. Iteración: Termina cuando no es posible crear un nuevo agrupamiento. Se selecciona X_n, el patrón más alejado de los agrupamientos existentes. Calculado el umbral de distancia, si la distancia entre X_n y el agrupamiento más cercano es mayor que este umbral, se crea un nuevo agrupamiento cuyo centro es X_n y se hace otra iteración.
3. Agrupación libre: Los patrones no seleccionados en el paso previo como centro de agrupamiento, se agrupan libremente utilizando la regla de mínima distancia y recalculando el centro después de cada asignación

Algoritmo de Batchelor y Wilkins

- Ejemplo

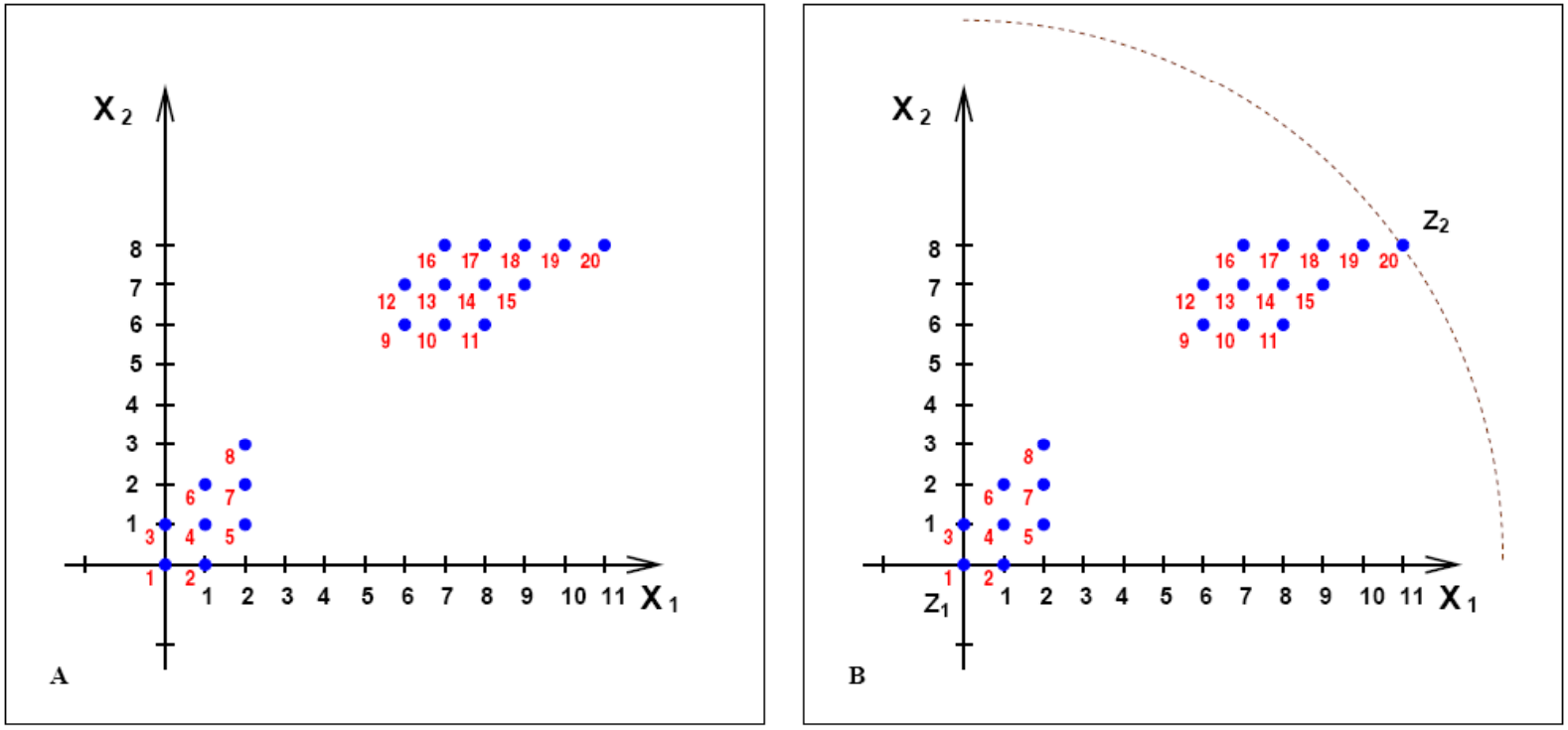


Figura 24: A) Patrones a agrupar. B) X_{20} es el patrón más alejado de Z_1

Algoritmo de Batchelor y Wilkins

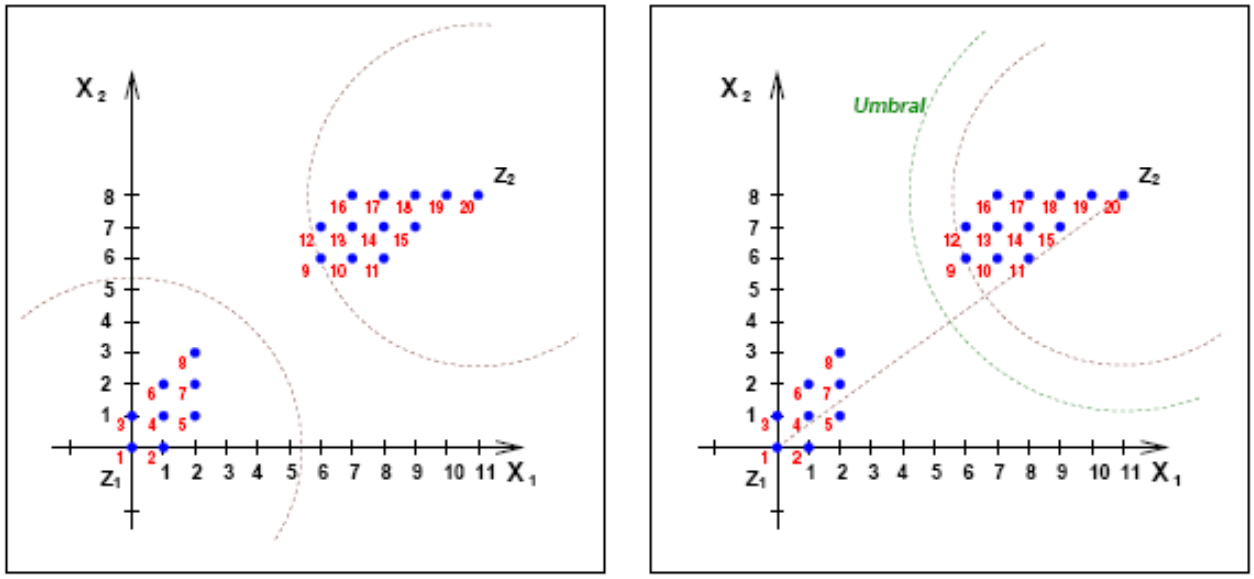


Figura 25: A) Cálculo de n. B) Cálculo del umbral

Algoritmo de Batchelor y Wilkins

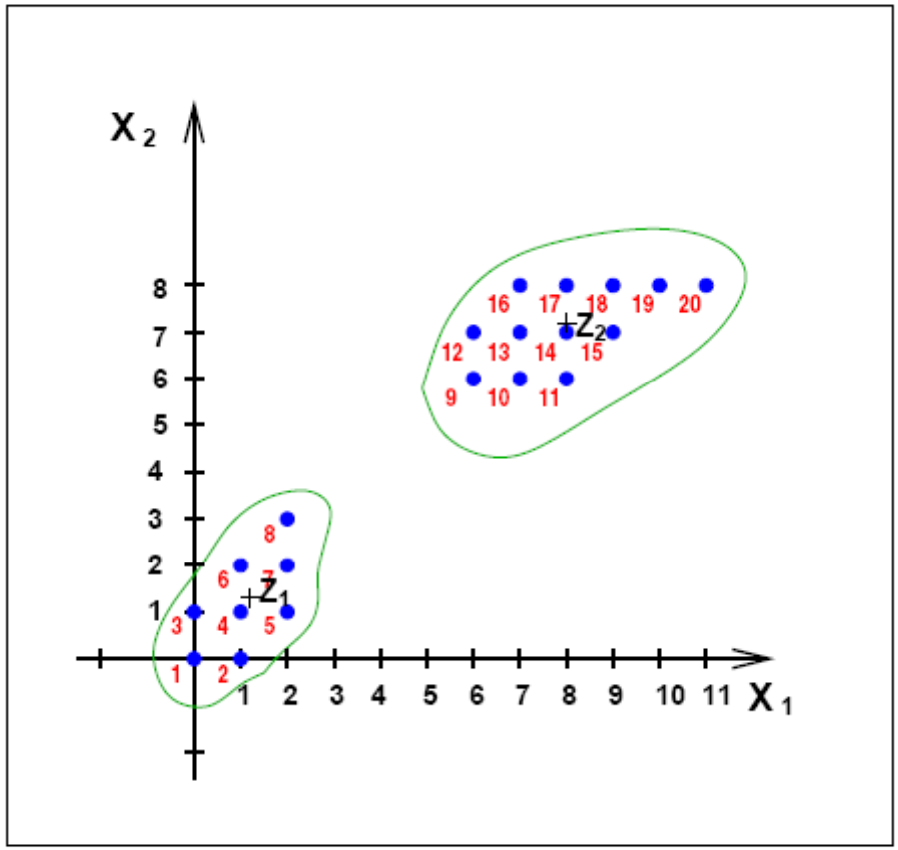


Figura 26: Resultado final

Algoritmos de Agrupamiento

- Número de clases desconocida
  - Método Adaptativo
  - Algoritmo de Batchelor y Wilkins
- Número de clases conocida
  - Algoritmo de las K-medias
  - Agrupamiento secuencial
  - Algoritmo ISODATA

Algoritmo de las K-medias

- En cada iteración se recalculan los centros de los agrupamientos.
- $S_i(t)$ es el conjunto de patrones asociados al agrupamiento S_i en la iteración t
- $Z_i(t)$ es el valor del centro en cada iteración

Algoritmo de las K-medias

Algoritmo:

1. Inicialización
   - Se inicializa arbitrariamente los centros de los K grupos
2. Asignación y actualización de centros
   - Cada patrón se asigna al grupo más cercano y se recalculan los centros en base a esta asignación
3. Convergencia
   - En el paso previo algunos patrones pueden cambiar de agrupamiento y en consecuencia los centros de éstos. Si no hay modificaciones se termina el agrupamiento

Algoritmo de las K-medias

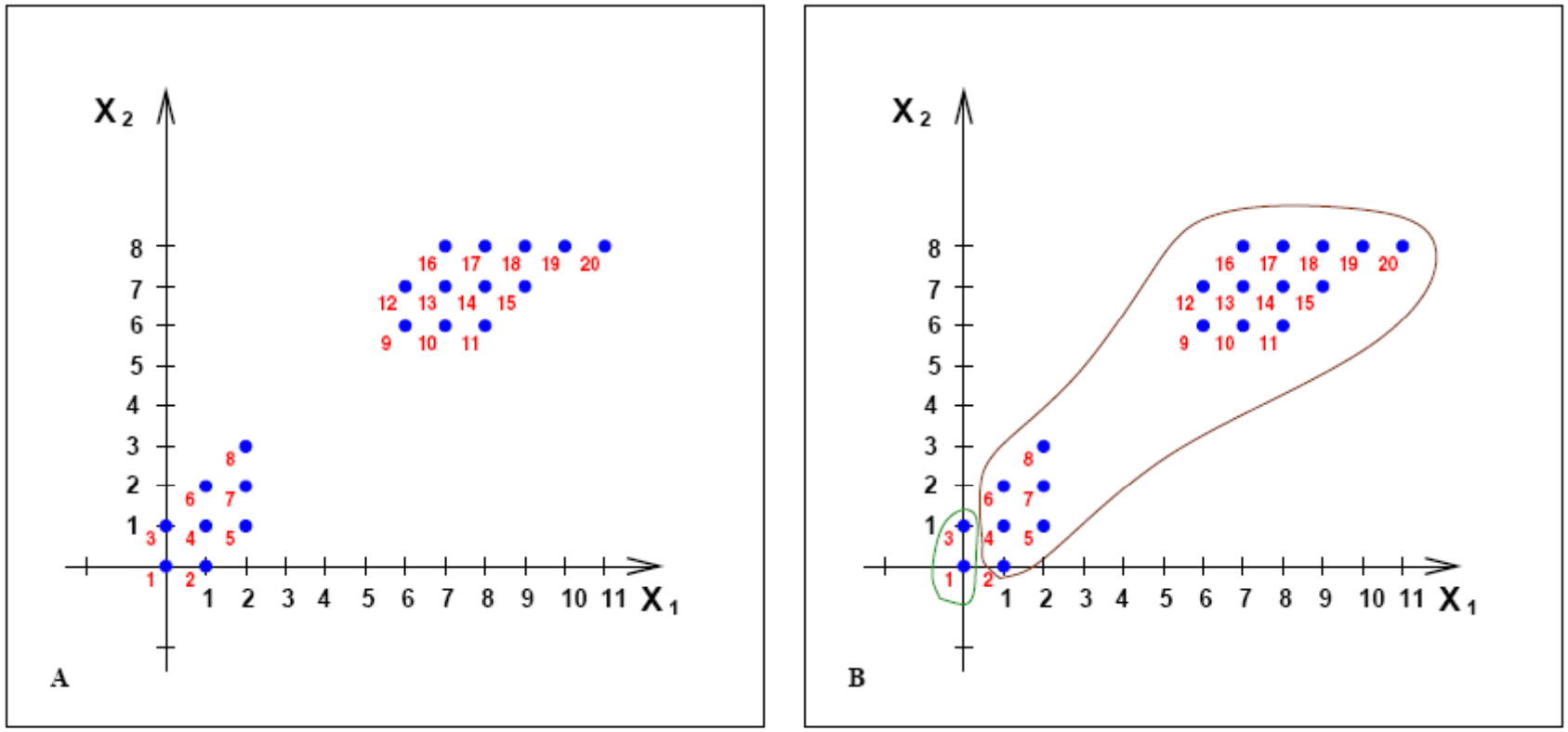


Figura 31: A) Situación inicial. B) Resultado de la primera asignación

Algoritmo de las K-medias

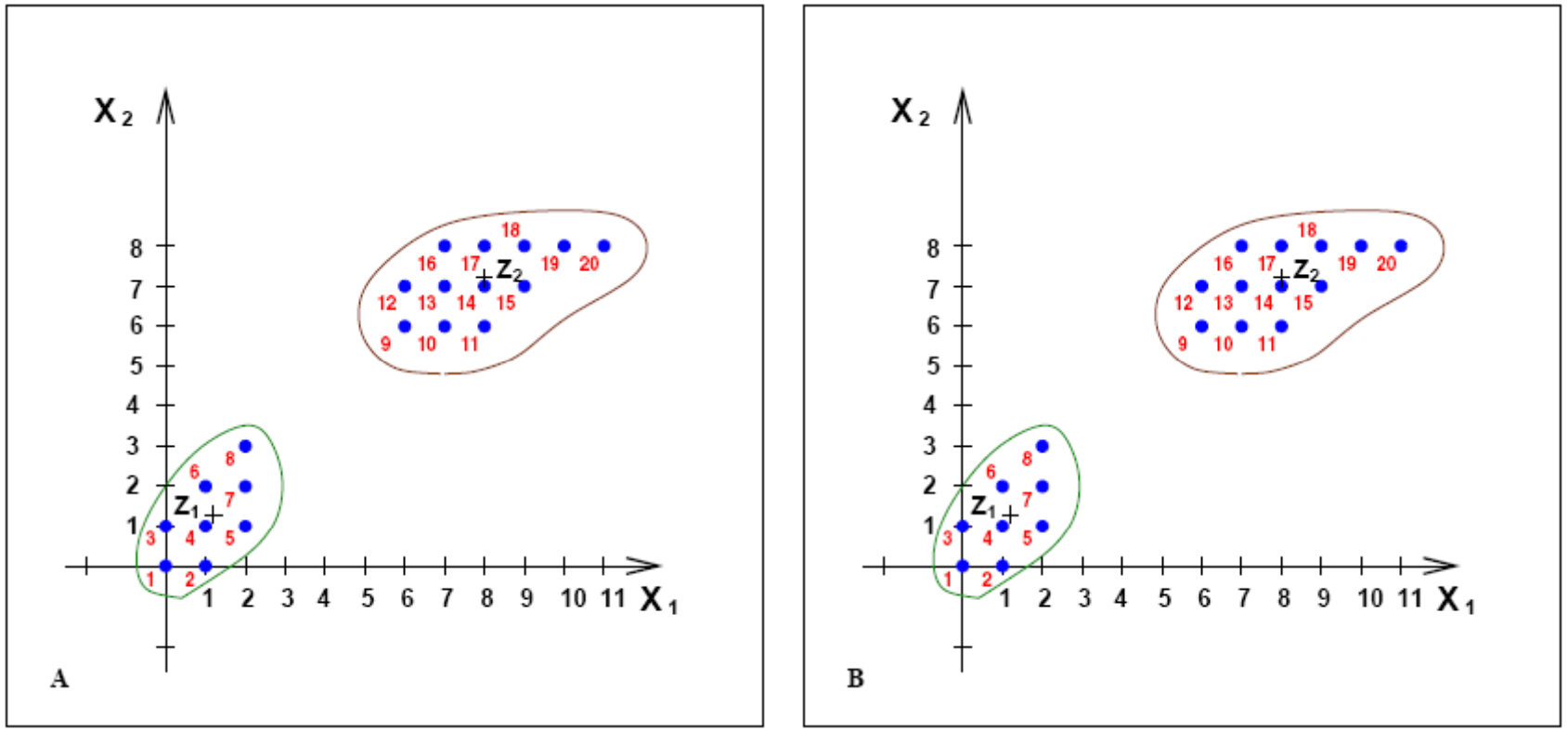


Figura 32: A) Segunda asignación. B) Tercera (y última) asignación

Algoritmos de Agrupamiento

- Número de clases desconocida
  - Método Adaptativo
  - Algoritmo de Batchelor y Wilkins
- Número de clases conocida
  - Algoritmo de las K-medias
  - Agrupamiento secuencial
  - Algoritmo ISODATA

Agrupamiento secuencial

Características:

1. El número de agrupamientos K se toma como máximo pudiendo devolver menor número de agrupamientos
2. Algoritmo incremental, crea agrupamientos a partir de un umbral de distancia
3. Fase de evaluación para reducir el número de agrupamientos en base a dos criterios:
   - El tamaño de los agrupamientos
   - La distancia entre los centros de los agrupamientos
4. Los patrones se procesan por lotes de manera que la evaluación de la partición se realiza al finalizar el lote

Agrupamiento secuencial

Algoritmo

1. Los patrones se procesan por lotes de longitud M
2. Durante el procesamiento de un lote los patrones se asignan al agrupamiento más cercano y se recalcula el centro. Si el agrupamiento más cercano está a una distancia superior al umbral R , se crea un nuevo agrupamiento. En esta fase el número de agrupamientos puede crecer por encima de K
3. Finalizado un lote se evalúa la partición con el objetivo de reducir el número de agrupamientos:
   a) Se mezclan parejas de agrupamientos que no disten más de un umbral C
   b) Se eliminan los que tengan pocos patrones
   c) Si no son aplicables ninguna de las anteriores, se aplica una mezcla forzosa hasta conseguir K agrupamientos

Agrupamiento secuencial

- Ejemplo: K=2, R=2, C=3.5, M=10, T=20

Orden de los patrones X_1, X_2, X_3, X_4, X_{10}, X_8,X_{20}, X_7, X_6, X_{11}, X_5, X_{16},X_{17}, X_{18}, X_{19}, X_9, X_{12}, X_{13},X_{15}, X_{14}

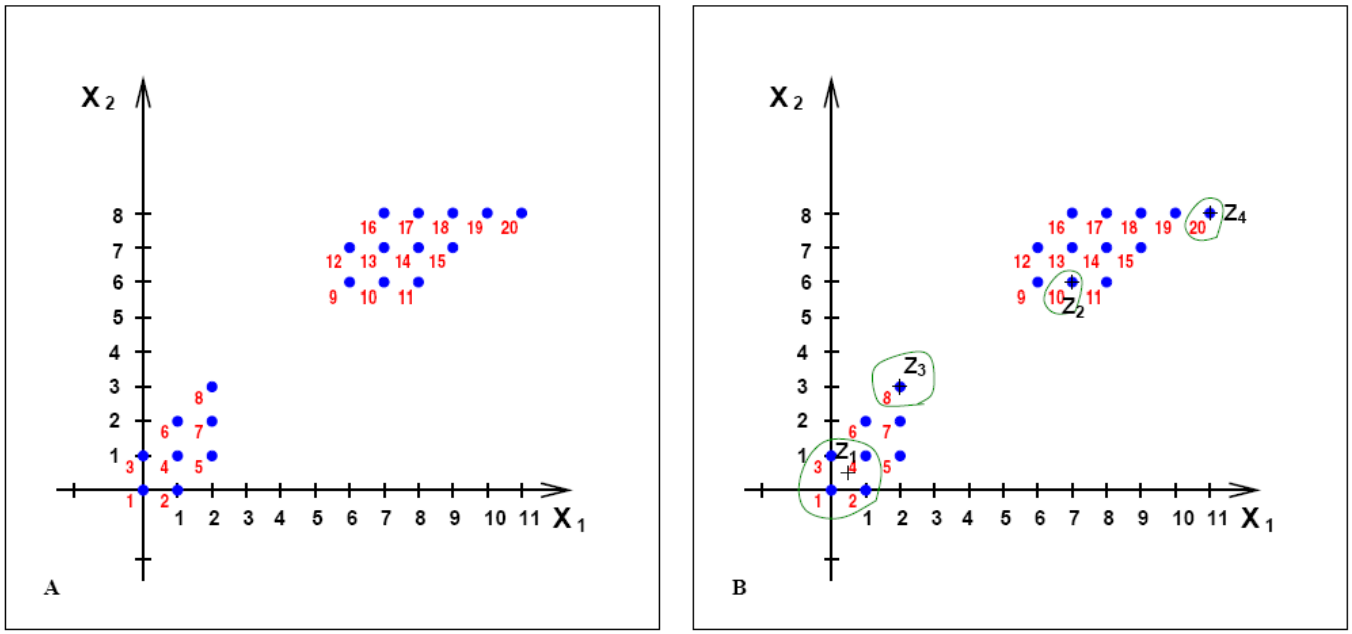


Figura 36: A) Situación inicial. B) Partición obtenida tras procesar los siete primeros patrones

Agrupamiento secuencial

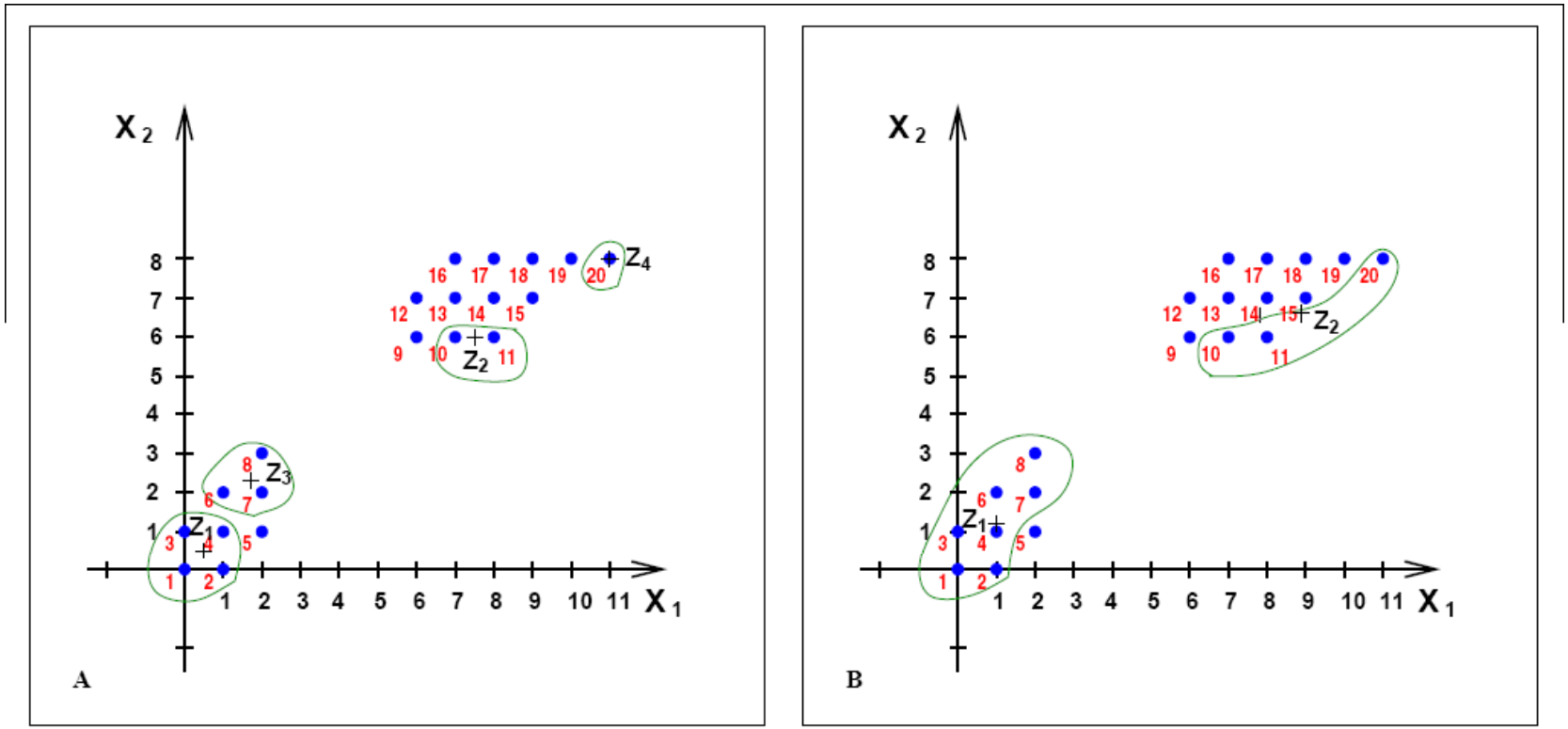


Figura 37: A) Estado al terminar el primer lote. B) Estado después de reducir el número de agrupamientos, antes de empezar el segundo lote.

Agrupamiento secuencial

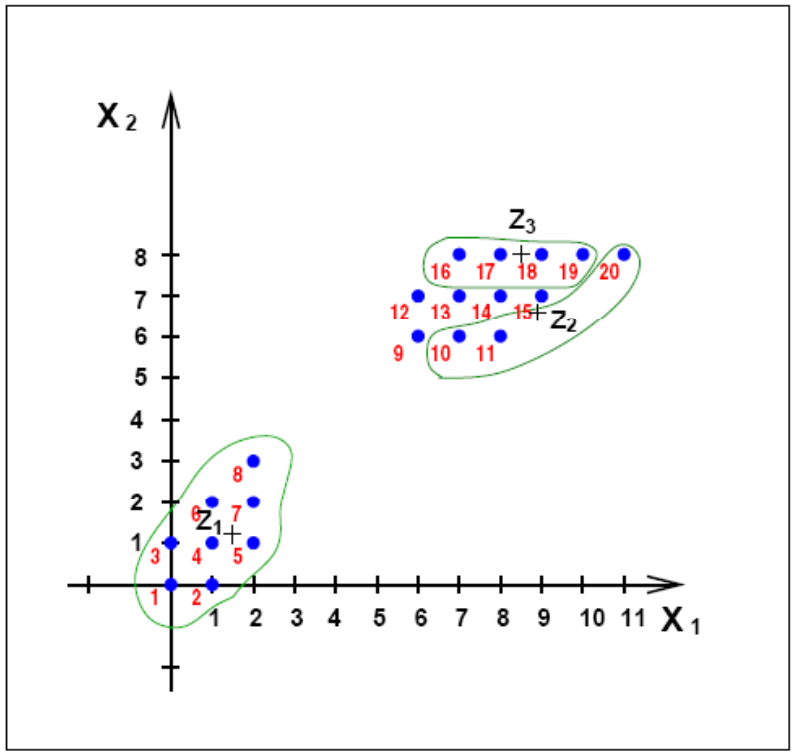


Figura 38: Estado tras procesar los quince primeros patrones

Agrupamiento secuencial

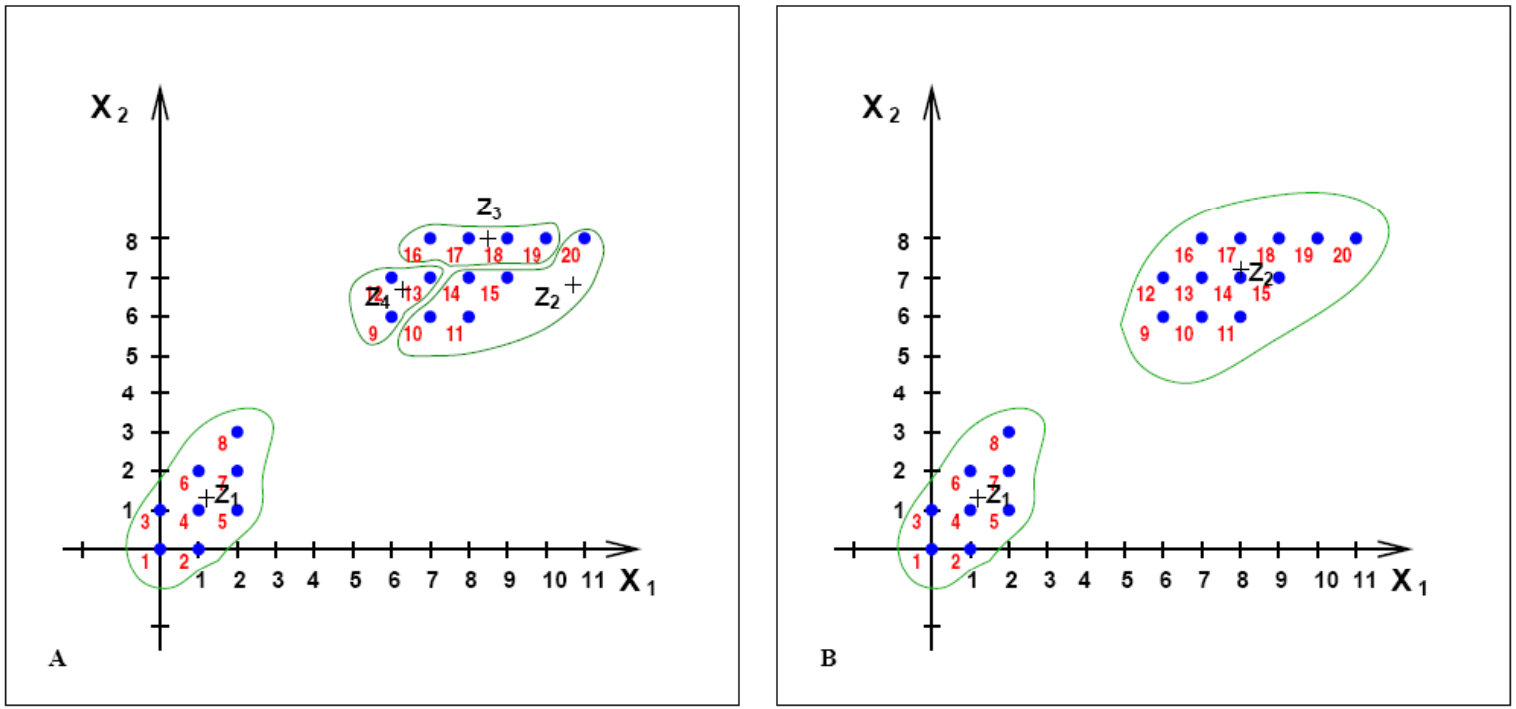


Figura 39: A) Estado al terminar el segundo lote. B) Estado final, después de reducir el número de agrupamientos.

Agrupamiento secuencial

- Ejemplo: Se procesan los primeros 15 patrones únicamente

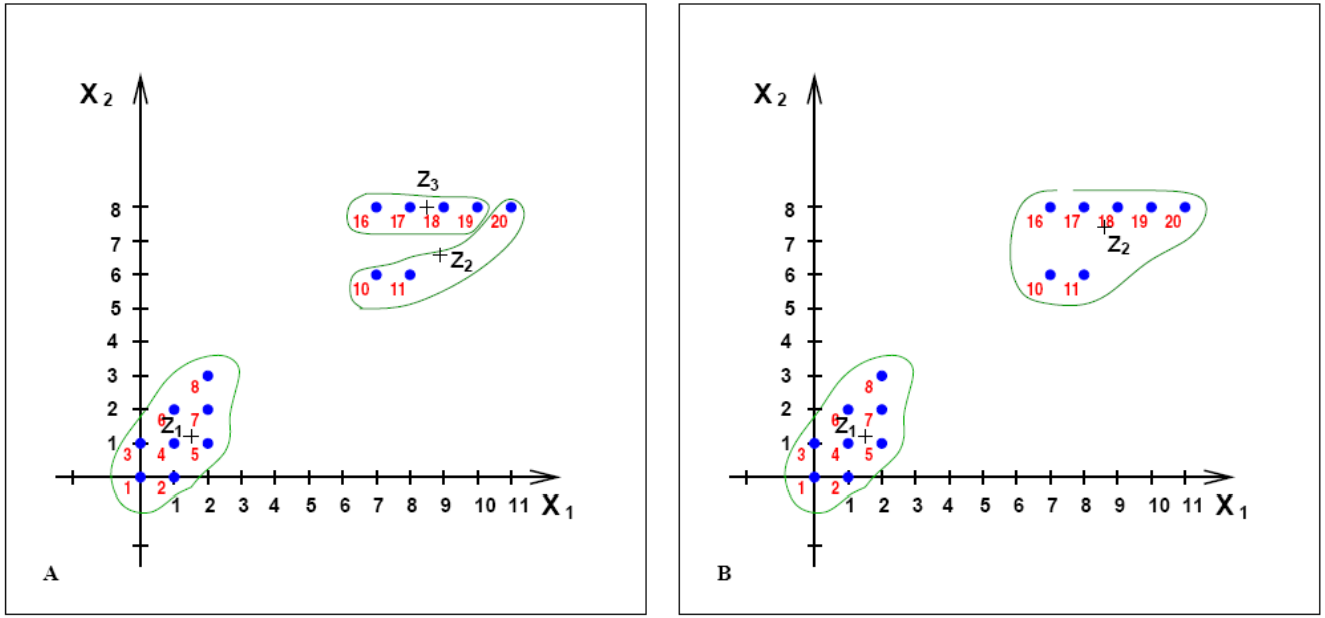


Figura 40: A) Estado tras procesar los quince primeros patrones. B) Estado final, después de reducir el número de agrupamientos.

Algoritmos de Agrupamiento

- Número de clases desconocida
  - Método Adaptativo
  - Algoritmo de Batchelor y Wilkins
- Número de clases conocida
  - Algoritmo de las K-medias
  - Agrupamiento secuencial
  - Algoritmo ISODATA

Algoritmo ISODATA

Iterative **S**elf-**O**rganizing **D**ata **A**nalysis **T**echniques

- Como en el K-medias los patrones se procesan repetidamente y en cada iteración se asignan al grupo más cercano
- Incorpora una serie de heurísticas con objeto de:
  - Eliminar agrupamientos poco numerosos
  - Mezclar agrupamientos cercanos
  - Dividir agrupamientos dispersos