

### 3. Clasificación no supervisada

El presente capítulo y el siguiente tratan de clasificación, es por ello que antes de abordar el tema específico de este capítulo, previamente se hará una introducción al tema de clasificación.

#### 3.1 Introducción

##### Importancia y propósito

Clasificar ha sido, y es hoy en día, un problema básico en un amplio espectro de disciplinas que se extiende de las ciencias básicas a la ingeniería. Dependiendo de la ciencia y del periodo histórico, el problema de clasificar lleva consigo su propia terminología: desde taxonomía al tan actual reconocimiento de patrones. Los conceptos aquí discutidos son de utilidad en una variedad de áreas de ingeniería de petróleo, en particular, cuando se necesita, condensar información, y reconocer patrones, como por ejemplo, en la identificación de: litologías a partir de registros, modelo de una prueba de pozo, patrones de flujo en flujo multifásico, y bloques y celdas en la construcción de modelos de yacimientos.

El propósito fundamental, común a todas las disciplinas, consiste en hacer una partición de un conjunto de objetos en categorías. Estas categorías, (o sus sinónimos: clases, conglomerados, grupos, etc.), se construyen de manera tal que un objeto en un grupo dado es similar, en algún sentido, a cualquier otro del mismo grupo; y objetos en distintos grupos tienden a ser diferentes.

Cada objeto es observado mediante un conjunto de variables cuantitativas que reflejan las cualidades fundamentales del mismo. Cada objeto tiene asociado entonces un conjunto de valores sobre un conjunto de  $p$  variables, que en lo sucesivo se llamará una observación. El conjunto de observaciones se agrupa en una matriz  $X$  de dimensión  $(n \times p)$ .

Luego, el proceso de clasificar, que se lleva a cabo sobre la matriz  $X$ , consiste en: dado un conjunto de  $n$  observaciones y sus características dadas por  $p$  variables, se requiere agruparlos basándose en las semejanzas que existan entre sí.

##### Metodologías para abordar la clasificación

Las metodologías de clasificación provienen fundamentalmente de dos fuentes: El análisis estadístico multivariado y el área de la inteligencia artificial llamada computación emergente. Los métodos pueden organizarse así:

- Análisis estadístico multivariado
  - Análisis de conglomerados (cluster)
  - Análisis discriminante
- Computación emergente
  - Redes neuronales
    - \* Perceptrón multicapa

- \* Mapas auto-organizativos
- Lógica difusa

Gran parte de la teoría estadística del análisis multivariado, que constituye el núcleo de los procesos clasificatorios fue desarrollada en la primera mitad de este siglo. Sin embargo, dadas las dificultades de cálculo, sólo podían abordarse pequeños problemas: limitados tanto en el número de observaciones como en el de variables que caracterizaban a los objetos. Los algoritmos de computación emergente, que no exigen conocimiento previo del tipo de distribución de probabilidad, han probado ser muy eficientes para abordar problemas de data compleja.

En las últimas décadas los algoritmos de clasificación se implementan eficientemente sobre un computador y proveen los resultados sin intervención humana. Sin embargo, en la mayoría de las aplicaciones tecnológicas, el procesamiento obtenido es sólo un instrumento de soporte en la toma de decisiones, y es el usuario que conduce el proceso de Data Mining quien deberá decidir, por ejemplo: ¿en cuántas categorías clasificar la población de objetos?; ¿debe la clasificación ser jerárquica?; ¿un objeto alejado estadísticamente de las clases existentes es el anuncio del descubrimiento de una nueva clase o debe forzársele a pertenecer a una de las clases existentes?

### **Extractores de características.**

Si bien la capacidad de cálculo de los actuales computadores permite resolver eficientemente gran parte de los problemas de clasificación no es menos cierto que cada vez la complejidad de los problemas de clasificación va en aumento: tanto en el número  $n$  de observaciones a clasificar como en la dimensión  $p$  del espacio de variables que definen el objeto.

### **El síndrome de la dimensionalidad.**

La mayoría de los algoritmos clasificatorios padecen del síndrome de la dimensionalidad: probada eficiencia para problemas de dimensión reducida pero se vuelven ineficientes en problemas de gran escala. Es así que para espacios donde la dimensión  $p$  es excesiva se vuelve indispensable reducir la dimensionalidad del mismo. Los procedimientos que llevan a cabo esa función se denominan extractores de características.

### **Propósito**

El objetivo fundamental de un extractor de características en procesos de clasificación es encontrar una transformación desde el espacio de dimensión  $p$  de las variables asociadas a cada observación en un espacio de dimensión inferior, denominado espacio de las características, que retenga de cada observación lo esencial de la información necesaria para el proceso de clasificación. Más precisamente: que el proceso clasificador de las observaciones en el espacio de la totalidad de las variables y en el espacio de las características conduzca a una división de las observaciones en las mismas clases o con diferencias insignificantes.

Obviamente, la terminología de espacio de las características obedece a que de las numerosas variables que representan la observación se extraen las características esenciales de las mismas.

Existen tres razones principales para aplicar un extractor de características. La primera, la complejidad computacional de los algoritmos de clasificación se reduce sensiblemente al trabajar sobre un espacio de dimensión inferior. La segunda, los métodos estadísticos de estimación se vuelven más confiables en un espacio de dimensión reducida. La tercera, la posibilidad de que la dimensión del espacio de las características no exceda de tres, para permitir una visualización gráfica de las clases en juego.

Los métodos para extraer características son los vistos en reducción de la dimensionalidad y otros que se presentarán en el capítulo siguiente.

### **Tipos de clasificación.**

Existe una división primaria en el concepto de clasificar:

- clasificación supervisada
- clasificación no supervisada.

La diferencia fundamental entre ambos métodos estriba en si se conoce o no la clase a la cual pertenece cada patrón (observación) de la data.

A continuación se aclararán estos conceptos.

La clasificación es supervisada si ya existe un conjunto de observaciones clasificadas en un conjunto de clases dado, y se conoce la clase a la que cada observación pertenece.

En la clasificación supervisada se distinguen dos fases fundamentales bien diferenciadas: la primera, consiste en el desarrollo o creación de una o varias reglas de decisión (diseño del clasificador), y la segunda, el proceso en sí de clasificación de nuevas observaciones.

En la primera fase, el conjunto cuyas clases ya están bien definidas se desglosa en un conjunto de entrenamiento y otro de validación. Se diseña el clasificador con el conjunto de entrenamiento y se observa su capacidad para clasificar con el conjunto de validación. En la segunda fase se procede a clasificar nuevas observaciones de las que se desconoce la clase a la que pertenecen.

La clasificación es no supervisada cuando se dispone de un conjunto de objetos (observaciones), donde se desconoce tanto el número de clases en que es razonable particionarlo así como a qué clase pertenece cada observación.

Este proceso de clasificación no supervisada, es significativamente más complejo que el de la supervisada ya que se desconocen las clases naturales, y dependerá de la habilidad para seleccionar:

- las características que representan al objeto (elección de las variables que constituyen una observación)
- la metodología de clasificación

### 3.2 Clasificación no supervisada

#### Definición

Este proceso de clasificación consiste en:

Agrupar un conjunto de  $n$  objetos, definidos por  $p$  variables, en  $c$  clases, donde en cada clase los elementos posean características afines y sean más similares entre sí que respecto a elementos pertenecientes a otras clases.

La similitud entre observaciones se establece en términos de distancias tal como se expondrá en esta sección. El número,  $c$ , de clases puede estar preestablecido o no, y depende del método elegido.

Varios son los propósitos que pueden conducir a este tipo de clasificación:

- Gráficar grupos afines, como es el caso de los dendrogramas de las taxonomías.
- Clasificar, simplemente, información abundante y compleja
- Hallar el número  $c$  de clases adecuado
- Encontrar subclases dentro de clases naturales
- Conceptualizar, interpretar los patrones analizando las causas intrínsecas de la formación de los mismos
- Hallar clases ocultas no previstas
- Preprocesar datos complejos con la finalidad de reducir la información a la aportada por los centros de las clases, para posteriormente realizar otros análisis con esta información simplificada, (Caras de Chernoff, por ejemplo).

Los métodos a ser considerados en este capítulo se desglosan en:

- Análisis de conglomerados
  - Directos
  - Jerárquicos
- Mapas autoorganizativos
  - Mapa de Kohonen
  - Mapa de Ultsch

#### 3.2.1 Formulación matemática

Toda la información disponible reside en la matriz  $X$  de dimensión  $(n \times p)$  de las observaciones. Esta información puede ser relevante tanto en el espacio de dimensión  $p$ ,  $\mathbb{R}^p$  de las variables, (espacio de los vectores filas), así como en el de dimensión  $n$ ,  $\mathbb{R}^n$ , de las

observaciones (vectores columnas).

Por ejemplo, la distancia entre dos observaciones (que es un concepto fundamental en clasificación no supervisada), se refiere a la distancia entre vectores fila y se determina en el espacio de dimensión  $p$ . Por el contrario, la correlación muestral entre las  $p$  variables - decisivo para la extracción de características, ya que dos variables altamente correlacionadas contienen casi la misma información y puede eliminarse una de ellas- se determina en el espacio de dimensión  $n$  de las columnas.

Se supondrá en lo que sigue que cada observación es un punto del espacio de dimensión  $p$  de las variables y que dado dos observaciones,  $X_i$  y  $X_j$ , la distancia entre ambas es la distancia Euclídea habitual.

### Objetivos duales en la clasificación

Existen dos objetivos duales en el proceso de obtener una clasificación óptima:

- Minimizar las desviaciones entre las observaciones que pertenecen al mismo grupo
- Maximizar las distancias entre los centros de los grupos

Se formulan como duales ya que es posible probar que basta lograr uno de los objetivos para que simultáneamente se logre el otro.

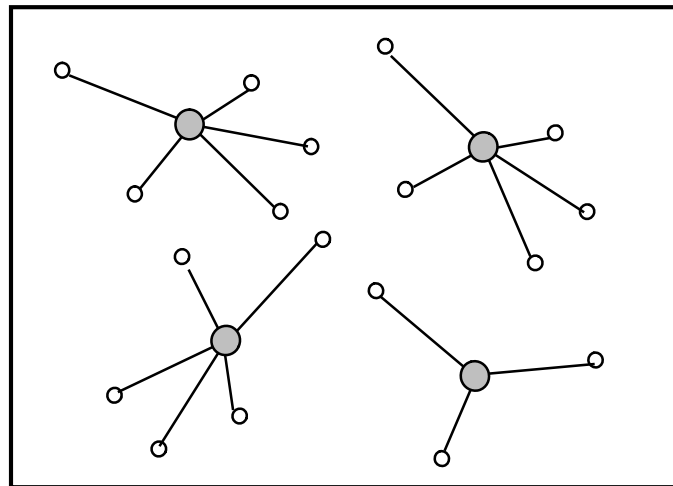


Figura 3.1. Dispersiones de las clases

### Definición

Se llamará  $SW_j$ , dispersión en la clase  $j$ , de  $N_j$  elementos, a la suma de las distancias al cuadrado de cada observación  $X_i$  al centro  $m_j$  de la clase ( $j$ ) que la contiene:

$$SW_j = \sum_{i=1}^{N_j} \|X_i - m_j\|^2 \quad (1)$$

Si  $m$  es el centro de la data, la dispersión total de la data está dada por:

$$ST = \sum_{i=1}^N \|X_i - m\|^2 \quad (2)$$

Uno de los objetivos duales es hacer que la dispersión sea mínima, pero de nada sirve que una clase este muy concentrada en torno a su centro y otras no. Es por ello que es necesario establecer una medida de concentración global de la totalidad de las clases,  $PW$ , que es la suma de las dispersiones,  $SW_j$ , de cada clase. Es esta medida la que debemos minimizar para optimizar la clasificación. Es posible redefinir ahora el objetivo de la clasificación no supervisada: fijado el número  $c$  de clases, distribuir las observaciones en  $c$  clases de modo de minimizar:

$$\min PW = \sum_{j=1}^c SW_j \quad (3)$$

La determinación de una clasificación no es nada trivial y es necesario complejos algoritmos para hallar simplemente soluciones satisfactorias.

Para tener una medida de bondad de ajuste de la clasificación que sea comparable con otras clasificaciones, se propone, el indicador:

$$R^2 = 1 - \frac{PW}{ST} \quad (4)$$

El indicador, que es análogo con el de los modelos lineales, cumple:  $0 \leq R^2 \leq 1$ . Si la clasificación es adecuada,  $PW$ , debe ser pequeño. De modo que cuanto mayor sea  $R^2$  mejor es la clasificación.

### 3.2.2 ¿Cuántas clasificaciones son posibles?

Anteriormente se expresó la dificultad de encontrar un algoritmo óptimo ¿a qué es debida esta dificultad, si para una partición dada el cálculo de  $PW$  es tan simple?. La razón es que el cálculo exhaustivo de  $PW$  para cada una de las particiones posibles se torna inabordable, aún para los computadores más eficientes del mundo, cuando inclusive los valores de  $n$  y  $c$  son relativamente pequeños.

El número de particiones de un conjunto de  $n$  elementos en  $c$  clases está dado por los números de Stirling de segunda clase. Por ejemplo:

n	c	participaciones
8	3	966
12	4	611.501
15	4	42.355.950
20	5	749.206.090.500

Vista la dificultad intrínseca de la optimización, se han propuesto diversos métodos para obtener soluciones razonables que se desglosan en:

- Análisis de conglomerados
  - Directos
  - Jerárquicos
- Mapas auto-organizativos

### 3.3 Análisis de conglomerados

El análisis de conglomerados, de extenso uso durante el siglo pasado en problemas de reconocimiento de patrones admite una clasificación primaria en:

- Métodos que clasifican a partir de la matriz de distancia entre todas las observaciones de la data. Entre estos métodos se cuentan los jerárquicos.
- Métodos directos, que sólo calculan distancias de las observaciones a posibles centros de las clases para luego modificar estos últimos sin necesidad de usar, las distancias entre las observaciones.

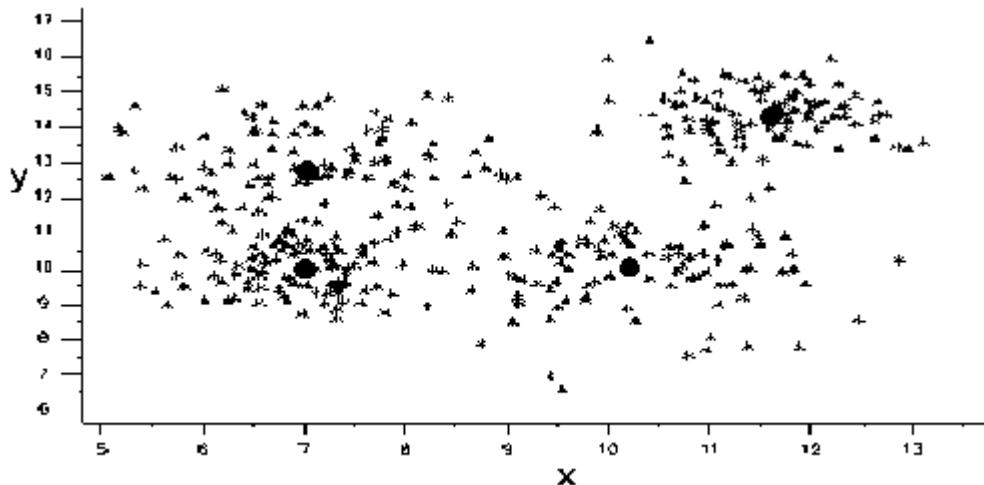


Figura 3.2. K-means: centros finales

### 3.3.1 Métodos directos

Los métodos directos se caracterizan por:

- Son iterativos
- Calculan las distancias de las observaciones a posibles centros de las clases, para luego modificar estos últimos siguiendo el criterio de optimización
- No hacen uso de las distancias entre los elementos
- El número de clases se fija de antemano

Usados principalmente cuando  $n$  es grande ( $n > 10000$ , por ejemplo).

#### Algoritmo K-means

El método directo más ampliamente usado es el algoritmo iterativo de evolución de los centros (k-means). Consta de las siguientes etapas:

1. Ubicación tentativa de los centros iniciales de las clases
2. Asignación de las observaciones a la clase más cercana
3. Determinación de los nuevos centros de las clases
4. Verificar si se cumple alguno de los criterios de finalización del algoritmo. En el caso de no satisfacerse el criterio de convergencia se vuelve a la etapa 2.

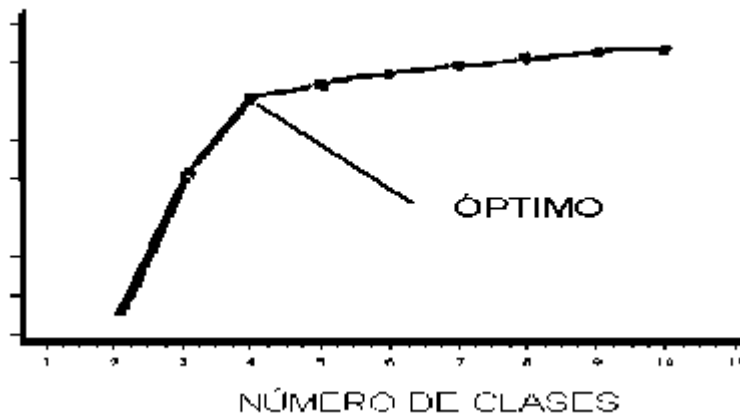


Figura 3.3. Evolución de R-cuadrado

Nótese que si las observaciones son 10.000 y las clases 10 en cada iteración se determinan 100.000 distancias; sin embargo, en un método jerárquico, donde se toman las distancias entre todas las observaciones, sería necesario calcular  $5000 \times 9999$  distancias!!

La figura 3.2 muestra los centros finales de una clasificación en 4 clases, de una data de 400 observaciones de dos variables, obtenida con el algoritmo FASTCLUS del paquete estadístico SAS. Se observa en la nube de puntos que una clasificación en tres clases podría

ser también válida.

### Determinación del número óptimo de clases

¿Cómo saber cuántas clases deben hallarse? Un criterio similar al existente para modelos lineales es estudiar el indicador  $R^2$ , (ecuación 3.4) en función del número de clases. Es obvio que  $R^2$  crece al aumentar el número de clases y que si  $c = n$  entonces  $R^2 = 1$ , luego no se busca el máximo de  $R^2$  sino analizar su tasa de cambio.

En la figura 3.3 se analiza la evolución de  $R^2$ , en el proceso de clasificar la misma data vista en la figura 3.2. Se observa que incrementar el número de clases más allá de 4 no redundará en una disminución significativa de la dispersión. Es cuatro (4) el número óptimo de clases.

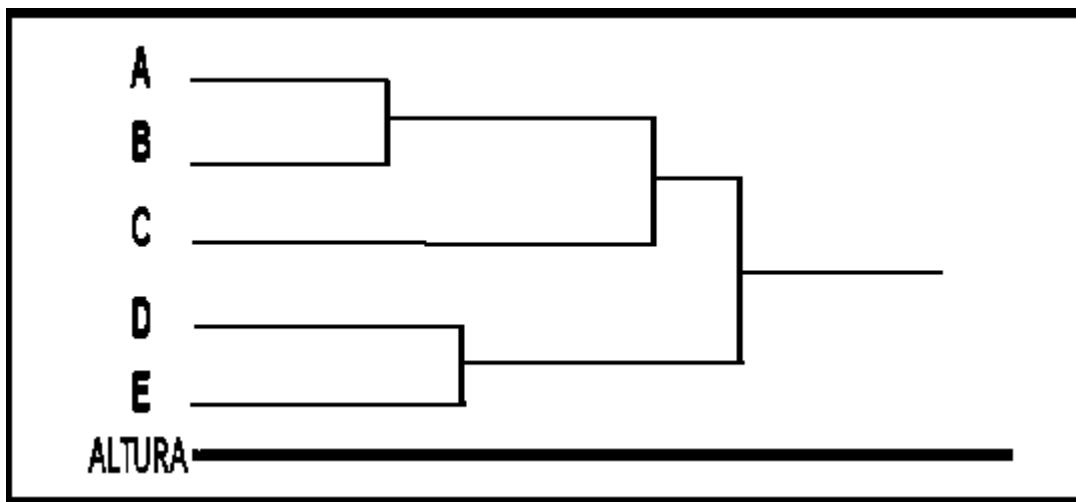


Figura 3.4. Clasificación jerárquica

### 3.3.2 Métodos jerárquicos

Los métodos jerárquicos crean una jerarquía entre las clases que se construyen a partir de las observaciones.

#### Propósito:

Dado un conjunto inicial donde cada elemento es una clase, crear un árbol jerárquico agrupando en cada etapa las dos clases ubicadas a mínima distancia, ésta indica la altura sobre el árbol.

La figura 3.4 es una jerarquía a partir de la matriz de distancias de 5 elementos. La primera clase que se constituye es  $\{a,b\}$  y luego  $\{d,e\}$ . La unión que sigue es entre la clase  $\{a,b\}$  y  $c$ ; debe existir un criterio para definir la distancia entre clases además de la distancia entre observaciones.

Los métodos jerárquicos se caracterizan por:

- Clasifican a partir de la matriz de distancia entre las observaciones
- No se fija el número de clases
- Se determina el número óptimo de clases a partir del árbol jerárquico
- Apropriados sólo si el tamaño del conjunto es pequeño, en cuyo caso son más eficientes que los métodos directos

La figura 3.5 es un árbol jerárquico de las distancias entre algunas ciudades en EEUU.

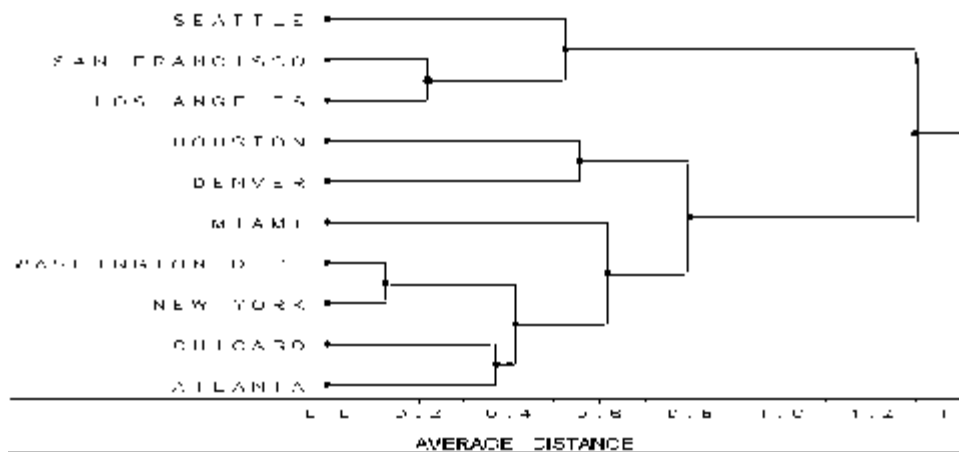


Figura 3.5. Árbol jerárquico

### Distancias

En el proceso de construcción de un árbol jerárquico surge la necesidad de definir la distancia entre clases además de la distancia entre elementos. Las distancias de uso frecuente son:

- Entre elementos
  - Euclídea
  - Estándar
  - Mahalanobis
- Entre clases
  - Distancia mínima
  - Distancia promedio ponderado
  - Distancia prototipo (centroide)

La distancia entre elementos estándar es la que corresponde a las variables originales estandarizadas, tal como se expuso en el tema de componentes principales. La distancia de Mahalanobis, incluye la información completa de la estructura de correlación; estimada su media, vector  $\mu$ , y su matriz de covarianza  $\Sigma$ , la distancia entre dos observaciones  $x$  e  $y$ , está dada por:  $(x - y)^T \Sigma^{-1} (x - y)$ .

En cuanto a las distancias entre clase, la distancia mínima es la que corresponde a la de las observaciones más cercanas de cada clase. La distancia más aplicada es la promedio ponderado que consiste en el promedio de todas las distancias entre elementos de cada clase. La distancia centroide es la distancia entre los centros de cada clase.

### 3.4 Mapas auto organizativos

Los mapas auto organizativos son muy eficiente en los proceso de clasificación de data compleja.

#### Propósito

Crear un mapa de Kohonen de la data original de modo que ésta se estructure en las neuronas del mapa. En cada neurona se agrupan elementos próximos que pueden conformar una clase por sí sola o en unión con clase vecinas.

El representante de la neurona en el espacio original se comporta como el centro de la clase aunque estrictamente no es obtenido de la misma forma que en los métodos directos.

#### Características del mapa

- Para que el mapa clasifique su tamaño debe ser reducido. Nótese que en reducción de la dimensionalidad se procedía del modo contrario.
- El número de clases suele ser inferior al tamaño del mapa ya que existen neuronas que no son ganadoras para ningún elemento de la data.
- Pueden detectarse observaciones aisladas (outliers). Éstas son atraídas en forma aislada por una neurona e inclusive puede pertenecer a una neurona se parada de las restantes en el mapa.
- Establece proximidad entre las clases y en consecuencia es posible unificar clases cercanas en el caso de que el número de clases sea excesivo.

#### Mapa de Ultsch

El mapa de Kohonen preserva topología pero no indica las distancias involucradas; es así que dos neuronas próximas en el mapa pueden distar mucho y ser muy inconveniente su unificación en una única clase. La clasificación mediante el mapa de Ultsch, (ver figura 3.6), soluciona el problema. Este mapa introduce nuevas neuronas entre las existentes que se colorean en una escala proporcional a las distancias entre las neuronas. Sobre este mapa es posible definir con más propiedad las clases a unificar.

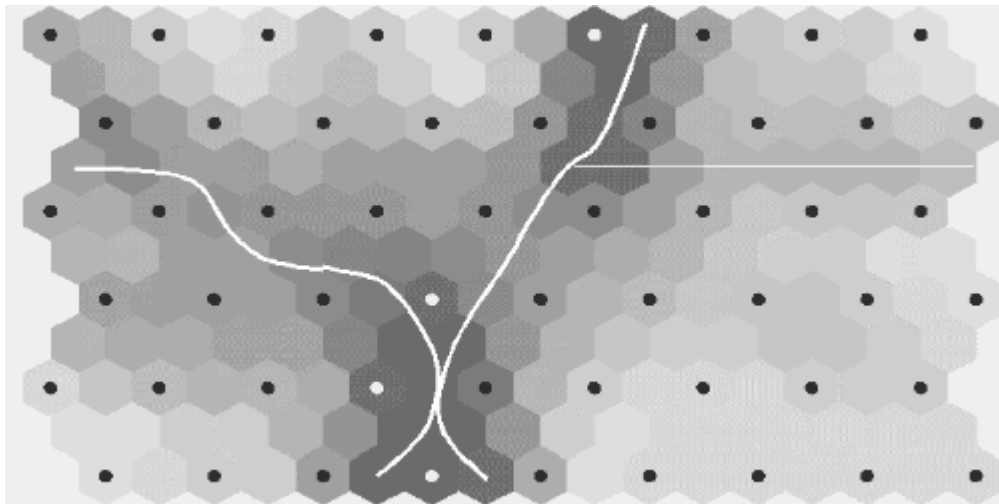


Figura 3.6. Mapa de Utsch