Universidad de Granada

Minería de Datos: Aprendizaje no Supervisado y detección de Anomalías
Apuntes Clustering

---

1. Introducción

A lo largo de estos apuntes, se analizará y profundizará en el **análisis cluster o aprendizaje no supervisado**, disciplina ampliamente estudiada dentro del campo de la minería de datos o **data mining**.

De manera informal, se puede definir un problema de aprendizaje no supervisado o clustering como aquel problema en el que, dado un conjunto de datos **no etiquetados**, se obtendrán una serie de grupos para estos datos. Las principales características de un problema de clustering o aprendizaje no supervisado son:

1. Los datos de entrada no están etiquetados y por tanto, no existe un resultado conocido.
2. El modelo es preparado deduciendo estructuras que se encuentran presentes en los datos y que es necesario descubrir mediante el análisis de los mismos.
3. Cualquier forma de descripción de los datos es igualmente buena si no existe ninguna evidencia que demuestre lo contrario.

De manera más formal, el aprendizaje no supervisado puede definirse de la siguiente forma:

Aprendizaje No Supervisado: *Dado un conjunto de datos x = (x1,x2,x3,..,xn), el objetivo es encontrar una función f que devuelve una descripción compacta del conjunto x. Es decir, el objetivo es encontrar una función f(x) que agrupe los elementos del conjunto de datos x en diferentes grupos.*

1.1. Conceptos Básicos

Una vez definido de manera formal y general un problema de análisis cluster o clustering, y antes de pasar a estudiar de manera detallada este tipo de problemas y las técnicas de agrupamiento que permiten abordarlos, es necesario conocer un conjunto de conceptos básicos y elementales que permitirán acercarnos más a los problemas que nos ocupan.

Julio Alberto López Gómez
Amparo Vila Miranda

Minería de Datos: Aprendizaje no Supervisado y detección de Anomalías
Apuntes Clustering

Agrupamiento: *Proceso de clasificar en grupos un conjunto de ítems sin tener una información previa acerca de su estructura.*

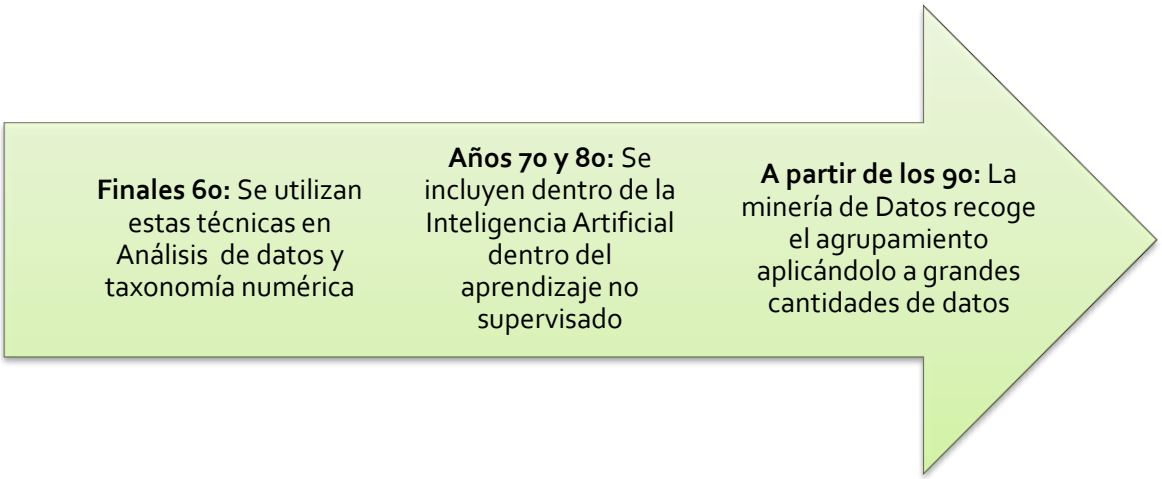
La definición anterior de agrupamiento, resume el proceso de clustering o aprendizaje no supervisado visto anteriormente, y que consistirá en dado un conjunto de datos cuya estructura se desconoce, clasificar dichos datos en grupos. Esta definición, más coloquial, da lugar a la siguiente definición, mucho más formal.

Agrupamiento: *El agrupamiento es una clasificación no supervisada de patrones (observaciones, datos o vectores de características) en grupos (clústers).*

Donde se introduce el concepto de **patrones** – crucial en minería de datos –. Estos patrones determinarán los elementos, observaciones o datos que pertenecen a cada uno de los grupos formados, así como el número de grupos.

De esta forma, los ítems o elementos a agrupar, los cuáles pueden estar formados por diversos atributos, también denominados variables, se distribuirán en cada uno de los grupos, según criterios que se verán en secciones posteriores. Las técnicas de agrupamiento o clustering se basan en la **minimización de la distancia intra-cluster**, es decir, los elementos que pertenecen a un mismo grupo deben de ser lo más parecidos posibles, y la **maximización de la distancia extra-cluster**, por la cual los elementos que pertenecen a dos clusters distintos deben ser lo más diferentes posibles.

Por último, y para concluir con esta sección de conceptos básicos acerca clustering y aprendizaje no supervisado, se muestra la siguiente línea del tiempo que resume los principales hitos temporales en cuanto al desarrollo y la evolución de las técnicas de agrupamiento se refiere.



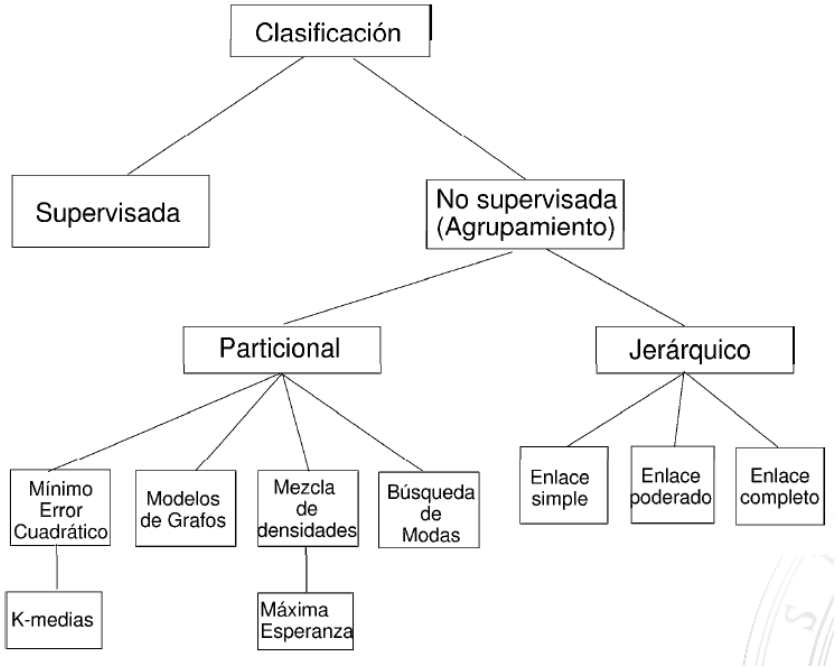
Julio Alberto López Gómez
Amparo Vila Miranda

1.2. Clasificación de las técnicas de clustering.

A la hora de estudiar las diferentes técnicas de agrupamiento, es necesario establecer una taxonomía o clasificación de la cual partir para abordar el estudio de los diferentes métodos de análisis cluster. A lo largo de las décadas, se han propuesto diferentes taxonomías, de acuerdo a los factores que influyen en las técnicas de agrupamiento, entre los que destacan:

1. ***La clase de problemas que se estén resolviendo.***
2. ***Las características intrínsecas de los datos de partida.***
3. ***Las medidas de semejanza que se estén utilizando.***

En base a estos factores, es posible establecer la siguiente clasificación de las principales técnicas de agrupamiento, las cuales se muestran en la siguiente imagen.



De donde en el segundo nivel del árbol distinguimos entre clasificación supervisada y no supervisada o agrupamiento:

Clasificación supervisada: En dicho tipo de clasificación los **datos** se encuentran **etiquetados**, por lo que la información de a qué grupo pertenece cada patrón es conocida, por lo que el objetivo de este tipo de clasificación es encontrar un conjunto de "criterios" que permitan clasificar un nuevo ítem. Generalmente, estos criterios serán un conjunto de reglas.

Clasificación no supervisada: O agrupamiento, se trata del ejemplo visto hasta el momento, donde **no se dispone de la información acerca de los grupos**, ni siquiera del número de grupos, sino que se trata de encontrar el mejor agrupamiento que reúna en el mismo grupo a los ítems más parecidos.

Dentro de la clasificación no supervisada, es posible hablar de agrupamiento particional y agrupamiento jerárquico.

Agrupamiento Particional: Es aquel tipo de agrupamiento en el que los **grupos** a obtener son **disjuntos** y cubren por completo el conjunto de ítems. Esto significa que un elemento no puede pertenecer a más de un grupo y que todos los elementos son clasificados.

Agrupamiento Jerárquico: En este caso, el resultado es una jerarquía de agrupamientos particionales anidados, de manera que cada grupo se divide en varios subgrupos, obteniendo una representación gráfica denominada **dendrograma**.

A su vez, dentro de las técnicas de **agrupamiento particional**, cabe destacar:

Mínimo Error Cuadrático: Se trata de aquellas técnicas que buscan la minimización del error cuadrático o del error cuadrático medio para realizar el agrupamiento. Una de las técnicas más utilizadas en este grupo es el algoritmo de las k-medias. Este método considera que los grupos deben ser cohesionados de manera que los ítems de un mismo grupo estén más cercanos entre sí y la distancia entre los grupos sea la mayor posible.

Modelos de Grafos: Son técnicas muy utilizadas cuando se considera que los datos están representados mediante un grafo donde los vértices son los ítems o patrones y las aristas las conexiones entre dichos ítems.

Mezcla de densidades: Técnica en la que se combinan zonas del espacio de diferente densidad para realizar el agrupamiento. En ellas se considera que un grupo es una región del espacio donde la densidad de ítems es muy alta y está rodeada de una región de baja densidad. Una de las técnicas más utilizadas dentro de este grupo es la denominada de máxima esperanza.

Búsqueda de modas: Técnica que utiliza el estadístico de la moda para realizar el agrupamiento.

Por otra parte, dentro de las técnicas de **agrupamiento jerárquico**, destacan:

Enlace Simple: Esta técnica mide la proximidad entre dos grupos calculando la distancia entre los ítems más próximos.

Enlace ponderado: Esta técnica mide la proximidad entre dos grupos calculando la media de las distancias entre objetos de ambos grupos, ponderando las distancias de alguno de ellos.

Enlace completo: Calcula la distancia entre los objetos más lejanos de dos grupos para medir la proximidad entre ellos.

Cabe destacar que todas las técnicas anteriormente nombradas parten de la hipótesis de **no-solapamiento** entre los conglomerados, de manera que los grupos son disjuntos entre sí. Cuando esta hipótesis se relaja, aparecen métodos que permiten el solapamiento o también llamados **no-exclusivos.**

Dentro de estas técnicas de agrupamiento no-exclusivo, destacan los **agrupamientos difusos**, los cuales permiten que un ítem pueda pertenecer a diversos grupos con un nivel de pertenencia a cada uno, modelando estos grupos como conjuntos difusos.

Por último, una vez vistas las técnicas más utilizadas de clustering, se definirán los conceptos de agrupamientos aglomerativos y divisivos que también servirán para clasificar las diferentes técnicas de clustering y para diseñar nuestros propios algoritmos de agrupamiento.

Agrupamiento Divisivo: *Se trata de aquel agrupamiento en el que inicialmente todos los ítems pertenecen a un mismo grupo. Conforme se realizan diferentes iteraciones, este grupo se va deshaciendo en otros subgrupos hasta que obtenemos un número de grupo constante y no observamos diferencias entre una iteración y otra.*

Agrupamiento aglomerativo: *Se trata de aquel agrupamiento en el que cada ítem es un grupo y se van construyendo nuevas soluciones uniendo grupos en otros más amplios.*

2. Proximidad, Distancia y Semejanzas

A la hora de comenzar a aplicar técnicas de clustering a un conjunto de datos, es necesario estudiar en primer lugar cómo viene representada la información de partida. Los datos con los que se trabajará pueden venir dados de dos formas:

1. Por medio de una **Dataset**: Es la forma más habitual en la que vienen dados los datos. En un dataset, cada uno de los elementos viene representado por una fila, mientras que las columnas representan diferentes atributos o variables de un mismo objeto. Un ejemplo de dataset se muestra en la siguiente imagen.

items\variables	V_1	V_2	$\cdots\cdots$	V_N
o_1	d_{11}	d_{12}	$\cdots\cdots$	d_{1N}
$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$
o_M	d_{M1}	d_{M2}	$\cdots\cdots$	d_{MN}

2. Por medio de una **Matriz de Proximidad:** Esta matriz de tamaño n x n, donde n es el número de objetos, representa en cada uno de sus elementos ij la distancia entre los elementos i y j. Esta distancia determina la similitud entre los elementos y viene dada por una **función de distancia**, las cuales se verán posteriormente. Generalmente, esta matriz se obtiene a partir del dataset, aunque en ocasiones, puede aparecer sin el conjunto de datos.

Julio Alberto López Gómez
Amparo Vila Miranda

Otro de los aspectos importantes a analizar antes de comenzar a aplicar técnicas de análisis de conglomerados, es el tipo de datos con el que se va a trabajar. Estos tipos de datos, se dividen en dos grandes grupos:

Datos o variables **Cuantitativas**: Son aquellos datos que quedan perfectamente definidos con un número. A su vez, las variables cuantitativas pueden ser:

1. **Discretas:** Aquellas medidas entre las cuáles no existen valores intermedios entre dos puntos. Ej: El número de televisiones que hay en un domicilio.
2. **Continuas:** En ellas sí que existen infinitos valores intermedios entre dos valores. La gran mayoría de las variables cuantitativas con las que más se trabaja son continuas. Ej: peso, altura…

Datos o variables **Cualitativas**: Son aquellas en las que sus valores no se expresan por medio de un valor numérico. En ellas se distinguen:

1. **Nominales:** Cuyo valor viene definido por una cadena de caracteres, como el color de los ojos o la nacionalidad de un individuo.
2. **Ordinales:** Sus valores vienen especificados por una cadena de caracteres que tiene un orden. En este tipo, destacan las variables ordinales binarias, como la presencia o no de enfermedad en un sujeto.

Proximidad.

Uno de los principales problemas que hay que abordar para comenzar el análisis de los datos de un dataset es construir una matriz de proximidad a partir de dicho dataset. Para ello, es necesario conocer los conceptos de **índice de proximidad** y **función de distancia**.

Índice de Proximidad: Sea un conjunto de n patrones, notados por $i, l, k.. \in I = \{1,2,\dots,n\}$ se dice que $d: I \times I \rightarrow R$ es un índice de proximidad si y sólo si verifica las siguientes propiedades:

1. Para medidas de disimilaridad o distancia: $\forall\, i \in I\ d(i,i) = 0$

2. Para medidas de similaridad: $\forall i \in I\ d(i,i) \geq \max_{k \in I} d(i,k)$

3. $d(i,k) = d(k,i) \forall i,k \in I$

4. $d(i,k) = d(k,i) \forall\, i,k \in I$

Donde la primera de estas propiedades especifica que la distancia de un elemento consigo mismo es cero. La segunda propiedad establece que la similitud entre un elemento y el mismo es siempre mayor o igual que la del elemento que más se parece a él.

Por último las últimas dos propiedades establecen, respectivamente, que la distancia entre dos puntos es simétrica y que la distancia entre dos elementos será siempre mayor o igual que cero.

Función de Distancia: Se dice que el índice de Proximidad d es una Función de Distancia si y sólo si verifica:

1. Las propiedades 1,3 y 4 de la definición anterior.
2. $d(i,k) \leq d(i,l) + d(l,k) \forall i,j,k \in I$

Esta última propiedad establece que la distancia entre dos elementos i y k será menor o igual siempre que la distancia entre i y un elemento cualquier l más la distancia entre ese elemento l y k – propiedad transitiva -.

La siguiente tabla muestra un conjunto de medidas de distancia conocidas y muy utilizadas en minería de datos, como son la distancia euclídea, Manhattan, Minkowski…

NOMBRE	EXPRESION
Euclídea o norma-l_2	$d_2(i,k) = [\sum_{j=1}^{m}(x_{ij} - x_{kj})^2]^{1/2} = [(\mathbf{x_i} - \mathbf{x_k})^{\mathbf{T}}(\mathbf{x_i} - \mathbf{x_k})]$
Manhattan o norma-l_1	$d_1(i,k) = \sum_{j=1}^{m} \|(x_{ij} - x_{kj}\|$
norma del supremo	$d_\infty(i,k) = sup_{j \in \{1,2..m\}} \|(x_{ij} - x_{kj}\|$
Minkowski o norma-l_p	$d_p(i,k) = \sum_{j=1}^{m} [\|x_{ij} - x_{kj}\|^p]^{1/p}$
Distancia de Mahalanobis	$d_M = [(\mathbf{x_i} - \mathbf{x_k})^{\mathbf{T}} \mathbf{\Sigma}^{-1} (\mathbf{x_i} - \mathbf{x_k})]$ Σ es la covarianza muestral o una matriz de covarianza intra-grup

Como se puede observar, la distancia de **Minkowski** es una generalización de todas las anteriores. No obstante, también existen funciones de distancia basdas en la distancia entre dos distribuciones de probabilidad o basadas en el coeficiente de correlación entre dos variables.

Entre todas las medidas que aporta la tabla anterior, la más intuitiva a la par que conocida es la **distancia euclídea**, que trabaja muy bien en casos en los que se tienen grupos "**compactos**" y "**aislados**".

Por otra parte, el principal problema de todas las denominadas medidas de Minkowski (aquellas que son una expresión concreta de la distancia de Minkowski) es que su uso da un gran peso a las variables con valores muy grandes. Sin embargo, este problema puede solucionarse, previa normalización de las variables del conjunto de datos. En cuanto a las variables continuas, otro problema habitual es la posible presencia de correlación entre ellas, lo cual se puede solucionar reduciendo el espacio o empleando la distancia de Mahalanobis.

Semejanza.

Otro de los principales factores asociados al estudio de un conjunto de datos a la hora de aplicar técnicas de agrupamiento, es el concepto de semejanza, que será descrito a continuación:

Semejanza: Dado un índice de proximidad s decimos que es una función de semejanza si y sólo si verifica:

1. $\forall i \in I \; d(i,i) = 1$
2. Las propiedades 3 y 4 de la definición de proximidad.

Julio Alberto López Gómez
Amparo Vila Miranda

Atendiendo a la definición anterior, es posible obtener un índice de semejanza a partir de una distancia, tal y como muestra la siguiente ecuación:

$$\forall i,k \in I\ s(i,k) = 1 - \left(\frac{d(i,k)}{D}\right) \text{ siendo } D = max_{i,k} d(i,k)$$

La gran mayoría de los índices de semejanza no basados en distancia, se definen para ítems cuyas variables son binarias. Si se consideran los ítems x_i y x_k formados por m variables binarias, se tiene:

- n_{IK} número de variables que toman el valor 1 en x_i y x_k
- n_{iK} número de variables que toman el valor 0 en x_i y x_k
- n_{iK} número de variables que toman el valor 0 en x_i y 1 en x_k
- n_{ik} número de variables que toman el valor 1 en x_i y 0 en x_k

Algunos de los índices de semejanza más utilizados son:

NOMBRE	EXPRESION
Indice de Jaccard	$\frac{n_{IK}}{n_{IK}+n_{iK}+n_{Ik}}$
Indice de acoplamiento simple	$\frac{n_{IK}+n_{ik}}{m}$
Indice de Russell	$\frac{n_{IK}}{m}$
Indice de Dice	$\frac{2n_{IK}}{2n_{IK}+n_{iK}+n_{Ik}}$
	$\frac{2(n_{IK}+n_{ij})}{m+n_{iK}+n_{Ik}}$
	$\frac{n_{IK}}{n_{IK}+2(n_{iK}+n_{Ik})}$
	$\frac{(n_{IK}+nik)}{m+n_{iK}+n_{Ik}}$

Otros índices de semejanza también conocidos son:

- **La medida del coseno:** Basada en la representación trigonométrica de cada documento como un vector de frecuencias de aparición de términos. Esta medida calcula el coseno del ángulo que forman ambos vectores.
  - **EJ:** Sean $t_1 = (t_{11} \dots t_{1D})$ y $t_2 = (t_{21} \dots t_{2D})$ dos vectores de documentos en un espacio d-dimensional, entonces.

$$\cos(t_1, t_2) = \frac{(t_1 \odot t_2)}{|t_1||t_2|}$$

Donde ⊙ representa el producto escalar y |.| el módulo. Por tanto:

$$cos(t_1, t_2) = \frac{\sum_{j=1}^{d} t_{1j} t_{2j}}{\sqrt{\sum_{j=1}^{d} t_{1j}^2} \sqrt{\sum_{j=1}^{d} t_{2j}^2}}$$

Con los conceptos vistos anteriormente, es necesario tener en cuenta una serie de **consideraciones**, a saber:

- **Tanto las medidas de distancias como las semejanzas se utilizarán para obtener la matriz de proximidad de un dataset, paso fundamental a la hora de comenzar un proceso de clustering.**
- **Cada uno de los enfoques anteriores corresponde a un tipo de variable.**
    - **Las distancias se utilizarán principalmente en variables continuas, aunque pueden usarse con valores enteros e incluso ordinales que sea posible categorizar como enteros.**
    - **Los índices de semejanza son muy útiles cuando se trabaja con factores binarios, además de poder utilizarse con variables nominales no ordinales transformándolas en un conjunto de factores binarios.**
    - **A la hora de tratar valores nominales, se pueden utilizar relaciones de semejanza previas difusas, las cuales suelen ser muy útiles en este tipo de problemas.**
    - **Es primordial tener mucho cuidado a la hora de mezclar diferentes enfoques cuando se tienen diferentes tipos de variables. En ese caso, habrá que establecer combinaciones de distancias y/o semejanzas convenientemente normalizadas.**
    - **A la hora de llevar a cabo un proceso de agrupamiento es crucial la preparación de los datos y la selección de la medida de distancia, siendo habitual que los resultados dependan mucho de estos factores y del tipo de problema. La selección, tanto de la distancia como la semejanza, es un proceso largo que requiere de diversos experimentos.**

3. Clustering Jerárquico

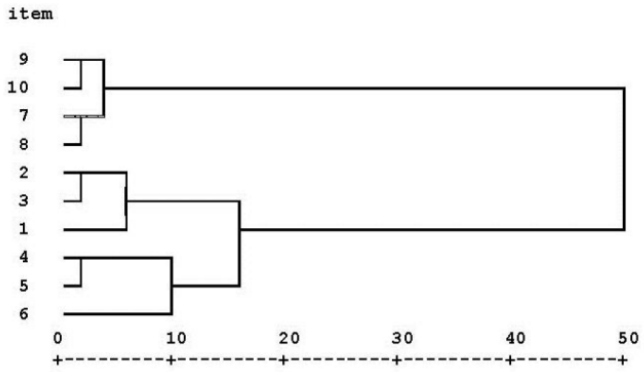
En el primer epígrafe de este documento, se definieron los métodos de clustering jerárquico y particional, haciendo hincapié en las diferencias entre uno y otro método. No obstante, en este apartado se profundizará especialmente en los métodos de agrupamiento jerárquico.

Clustering Jerárquico: Método de agrupamiento que parte de un único grupo formado por todos los elementos. A lo largo de la aplicación del algoritmo de agrupamiento, dicho grupo se irá particionando, de manera que mediante particiones "anidadas" se obtiene un número final de grupos y una clasificación de los diferentes ítems en cada uno de ellos.

Las características más notables de este tipo de algoritmos son:

- Consiste en una sucesión de particiones sucesivas o "*anidadas*".
- Cada grupo de ítems que pertenece a una partición está **totalmente incluido** en algún grupo de la partición siguiente.
- El resultado de la aplicación de las técnicas de agrupamiento jerárquico se puede representar de manera gráfica en un **dendrograma**, el cual representa cómo se van uniendo los distintos patrones en grupos.
- Mediante procesos algorítmicos y la **matriz de distancia** se obtiene el criterio de unión de grupos.

Un ejemplo de un dendrograma puede verse en la siguiente imagen, donde como se puede observar, aparecen diferentes particiones, unas dentro de otras que dan lugar, finalmente a nueve grupos.



Técnicas de Clustering Jerárquico.

La mayoría de los algoritmos de clustering jerárquico son de tipo **aglomerativo**. En este tipo de algoritmos se parte de un conjunto de grupos – tantos como elementos – donde cada elemento constituye un grupo y tras sucesivas iteraciones, estos grupos se van uniendo entre sí. No obstante, existen también algoritmos de clustering jerárquico que parten de un único grupo formado por todos los elementos que se va particionando en nuevos grupos. Estos enfoques, se vieron en el apartado de introducción y se denominan, respectivamente, enfoque **aglomerativo** y **divisivo**.

El enfoque utilizado por el clustering jerárquico se trata de un **enfoque basado en grafos**, donde cada ítem del conjunto de datos se considera un vértice del grafo y se van generando particiones, conectando aquellos vértices que se encuentran más cercanos, o que son más similares, de acuerdo a la medida de distancia seleccionada.

En este enfoque, aparecen dos formas de realizar el agrupamiento:

- **Agrupamiento de enlace simple o Single-Link Clustering**: En él los grupos se obtienen buscando las **componentes conexas** del grafo. El algoritmo termina cuando todos los vértices están conectados.
- **Agrupamiento de enlace completo o Complete-Link Clustering:** El algoritmo busca los subgrafos completamente conectados o **cliques**.

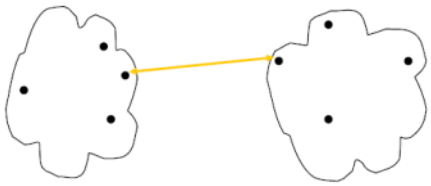
Uno de los algoritmos más conocidos y utilizados de agrupamiento jerárquico es el **Algoritmo de Jhonson**, que parte de la matriz de distancia de un conjunto de datos. Dicho algoritmo, trabaja de la siguiente forma.

1.El algoritmo realiza diversas **transformaciones sobre la matriz distancia**, reduciendo la dimensión de la misma siempre que se forme un nuevo grupo.

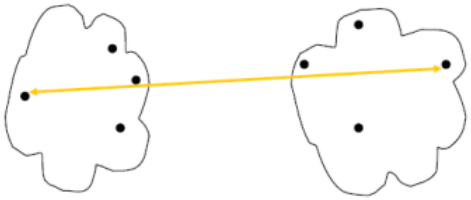
2. El algoritmo trabaja con una **matriz de distancia entre grupos**, la cual se va calculando iterativamente a partir de la matriz de la etapa anterior.

3. La distancia entre los grupos se puede calcular de distintas formas y dependiendo del cálculo de dicha distancia, aparecen diferentes formas de agrupamiento. Las diferentes medidas de distancia entre grupos o entre clusters son:

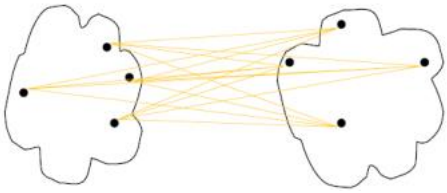
3.1. Enlace simple (Single Link): Es la mínima distancia entre un elemento de un cluster y un elemento de otro. Más formalmente, $d(C_i, C_j) = \min\{d(x_{ip}, x_{jq})\}$.



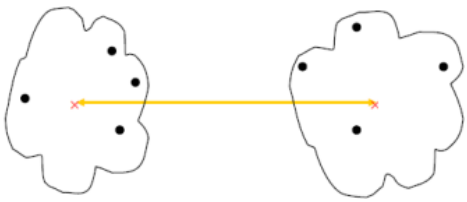
3.2. Enlace Completo (Complete link): Es la mayor distancia entre un elemento de un cluster y un elemento de otro. Más formalmente, $d(C_i, C_j) = \max\{d(x_{ip}, x_{jq})\}$.



3.3. Enlace ponderado o medio (Average Link): Es la distancia promedio entre los elementos de un cluster y los elementos de otro. Más formalmente, $d(C_i, C_j) = \text{avg}\{d(x_{ip}, x_{jq})\}$.



3.4. Enlace entre centroides (Centroid Link): Es la distancia entre el centroide de un cluster y el centroide del otro, siendo el centroide el elemento medio de un cluster.



Una de las técnicas de agrupamiento jerárquico más conocidas y extendidas es el algoritmo de Johnson, cuyo pseudocódigo se muestra a continuación.

Algoritmo de Johnson. Forma general de Lance y William

1.Sean $m = 0$, $D_m = D$, matriz de distancia de partida, $C_m = \{\{1\}, \dots, \{n\}\}$ el agrupamiento inicial, $L(m) = 0$ el nivel al cual se hace este agrupamiento.

2. Sean R y S aquellos grupos de C_m que tienen distancia mínima:

$L(m+1) = D_m(R,S)$

Formar un nuevo grupo $K = R \cup S$. Hacer

$C_{m+1} = C_m \cup (R \cup S) - R - S$ y transformar la matriz D_m de la siguiente manera:

Eliminar la fila y columna de S y asignar fila y columna de R a K.

Para todo $T \in C_m; C_m \neq K$ hacer

$$D_{m+1}(K,T) = a(R)D_m(R,T) + a(S)D_m(S,T) + bD_m(R,S) + c|D_m(R,T) - D_m(S,T)$$

3. Hacer $m = m + 1$

4. Si se han unido todo los ítems parar, en caso contrario ir a 2

Por último, una vez vistas las principales técnicas de clustering o agrupamiento jerárquico, y para comprobar las fortalezas y debilidades de cada una de ellas, este apartado concluirá con un ejemplo práctico que resumirá de manera gráfica y numérica los resultados de las técnicas de agrupamiento jerárquico.

En las transparencias de la asignatura, es posible ver un ejercicio donde se aplican las técnicas vistas anteriormente y se calcula su eficiencia, además de ilustrar el funcionamiento de cada una de ellas mediante el dendrograma resultante. En este caso, se plantea el siguiente ejemplo.

EJEMPLO: Dado un data sets de cinco objetos caracterizados por un único rasgo o atributo, se asumirá que existen dos clusters $C_1 = \{a, b\}$ y $C_2 = \{c, d, e\}$. Se pide:

	a	b	c	d	e
Feature	1	2	4	5	6

1. **Calcular la matriz de distancia**
2. **Calcular las distancias entre los clusters, utilizando los métodos de Single, Complete and Average Link.**

En primer lugar, se calculará la matriz de distancia. Para ello, tan solo es necesario fijarse en la tabla que aporta el enunciado. Al existir solo una característica, la distancia entre dos elementos estará simplemente en la diferencia entre sus valores en la variable o atributo que los identifica. En caso de que la tabla incluyera

más de una variable, sería necesario plantearse qué tipo de medida de distancia utilizar para ejecutar el algoritmo de agrupamiento.

Con este apunte, la matriz de distancia para los elementos del dataset es:

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

c

Por otra parte, se procederá a calcular la distancia entre los clusters de acuerdo a las expresiones dadas en la definición de Single Link, Complete Link, and Average Link. Los resultados de este experimento se muestran a continuación:

Single link

$$dist(C_1,C_2) = \min\{d(a,c), d(a,d), d(a,e), d(b,c), d(b,d), d(b,e)\}$$
$$= \min\{3, 4, 5, 2, 3, 4\} = 2$$

Complete link

$$dist(C_1,C_2) = \max\{d(a,c), d(a,d), d(a,e), d(b,c), d(b,d), d(b,e)\}$$
$$= \max\{3, 4, 5, 2, 3, 4\} = 5$$

Average

$$dist(C_1,C_2) = \frac{d(a,c)+d(a,d)+d(a,e)+d(b,c)+d(b,d)+d(b,e)}{6}$$
$$= \frac{3+4+5+2+3+4}{6} = \frac{21}{6} = 3.5$$

Las principales diferencias entre los tres tipos de enlace, vienen dadas en la siguiente tabla – resumen de estos tres métodos.

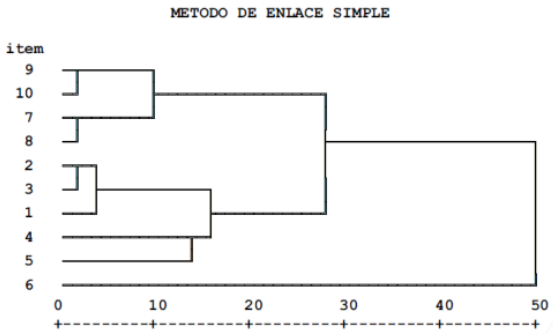
Método	Fortalezas	Debilidades
Single Link	Maneja formas no-elípticas	Sensible al ruido y outliers
Complete Link	Poco sensible al ruido y outliers	Tiende a dividir grandes clusters
Average Link	Poco sensible al ruido y outliers	Sesgada hacia ciertos clusters

Por su parte, en el ejemplo que se muestra en las transparencias de la asignatura, es posible ver la comparativa entre las técnicas citadas anteriormente y estudiadas en el ejemplo aplicadas sobre el algoritmo de Johnson, así como el dendrograma que generan. Dicha tabla comparativa y las imágenes, se adjuntan a continuación.

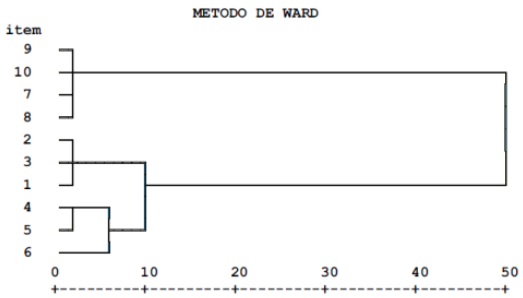
METODO	a(R)	a(S)
Enlace Simple	$1/2$	$1/2$
Enlace completo	$1/2$	$1/2$
Media de grupos	n_R/n_K	n_S/n_K
Centroide	$n_R/(n_R+n_T)$	$n_S/(n_s+n_T)$
Método de Ward	$(n_s+n_T)/(n_K+n_T)$	$(n_R+n_T)/(n_K+n_T$

METODO	b	c
Enlace Simple	0	$-1/2$
Enlace completo	0	$1/2$
Media de grupos	0	0
Centroide	$-(n_R n_S)/n_K^2$	0

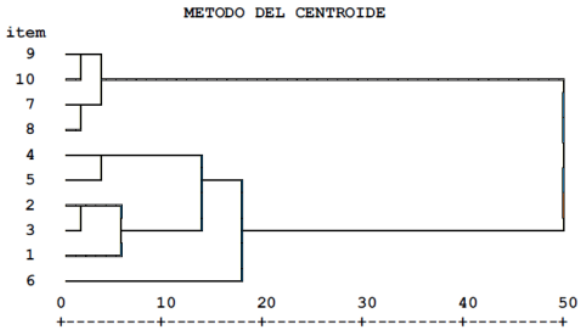
Dendrograma para el método de **Single Link**:



Dendrograma para el método de **WARD**: Este método define el criterio para mezclar dos clusters en función del valor óptimo de una función objetivo elegida por el investigador.



Dendrograma para el método del **Centroide**:



4. Clustering particional

Una vez vistas las técnicas de agrupamiento o clustering jerárquico, este apartado ahondará en el estudio de las técnicas de clustering particional. Al hilo de lo que se comentó en el primer apartado del presente documento, a continuación se definirá el clustering particional para, posteriormente, ver dos ejemplos de algoritmos de clustering particional, como son el método de las **k-medias** y el **DBSCAN**.

Clustering Particional: Dados n ítems representados en un espacio d-dimensional donde se ha definido una distancia, el objetivo consiste en determinar una partición de los mismos ítems en K subconjuntos o grupos tales que los ítems situados en un grupo se parezcan más entre sí que al resto de los situados en grupos diferentes (minimización de la distancia intra-cluster, maximización de la distancia extra-cluster).

A esta definición, cabe añadirle dos aspectos importantes:

1. El número K de grupos a generar puede estar definido a priori o no.
2. La similitud, coherencia o distancia entre un cluster y otro o entre un conjunto de clusters se mide mediante diversos criterios.

Atendiendo a esta definición, es posible definir e identificar una serie de **criterios globales** para todo algoritmo de clustering particional, así como un conjunto de **criterios locales** para todos ellos.

Criterios Globales:

- Los algoritmos de clustering particional suponen que cada grupo está representado por un prototipo (**Centroide**) y asignan cada ítem al grupo cuyo prototipo esté más cercano.
- Para ello se utilizan medidas de coherencia basadas en la distancia de cada ítem a los diferentes prototipos y en base a la distancia seleccionada se obtienen distintas medidas.
    - Para **atributos continuos**, el prototipo o **centroide** de un grupo suele obtenerse como la **media** de los atributos de los ítems que integran dicho grupo.

- En el caso de **variables categóricas**, el **centroide** se selecciona mediante la selección del ítem más representativo del grupo, denominado **medioide**.

Criterios Locales:

- Los algoritmos de Clustering particional forman los grupos utilizando la **estructura local** de los datos.
- El anterior principio es muy utilizado en métodos basados en la identificación de **regiones de alta densidad** de puntos o en aquellos métodos que asignan al mismo grupo un patrón y sus k-vecinos más cercanos.
- Uno de los métodos más conocidos en este contexto ,es **DBSCAN**.

4.1. El algoritmo de las k-medias

Dentro de los algoritmos de Clustering Particional, el algoritmo de las **k-medias** es, si no el más conocido y extendido, uno de los algoritmos más ilustrativos de esta categoría. El algoritmo de las k-medias ofrece un enfoque particional basado en **centroides**, donde una vez especificado el número **k** de clusters, se asignará cada ítem del conjunto de datos a aquel grupo cuyo centroide minimice la distancia entre el elemento y el centroide. De manera informal, el algoritmo de las k-medias puede describirse de la siguiente forma.

Algoritmo de las k-medias.

Parámetros Iniciales:

El número de grupos K y K centroides iniciales.

Proceso Básico:

1. Se asigna cada punto a su centroide más cercano. Se obtienen así los grupos iniciales.

2.A partir de estos grupos se recalculan los centroides y se hace una nueva reasignación.

3.El proceso se vuelve a repetir hasta que los centroides no cambian.

De manera más formal, el algoritmo podría ser descrito de la siguiente forma:

Algoritmo de las k-medias.

Seleccionar K puntos como centroides iniciales

Repetir

(re) asignar cada punto a su c entroide más cercano

(re) calcular el centroide de cada cluster

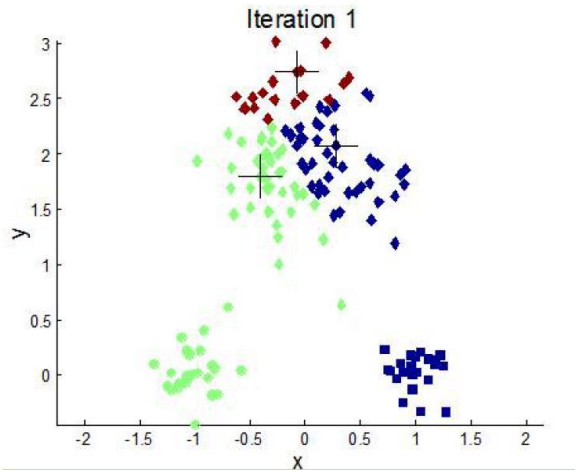
Hasta que los centroides no cambian (Criterio de Parada)

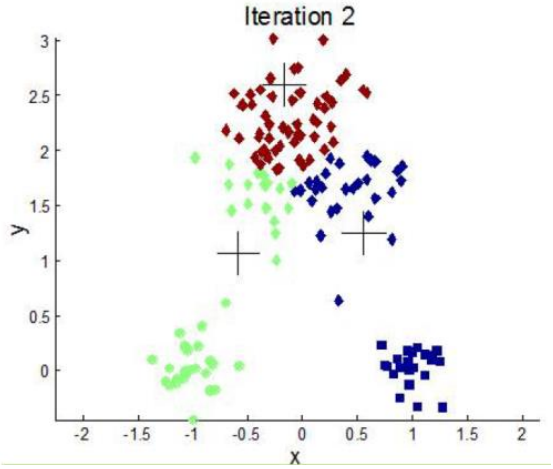
Resultado: Una partición $x_1, x_2, \ldots, x_n$ en K clusters determinados a priori.

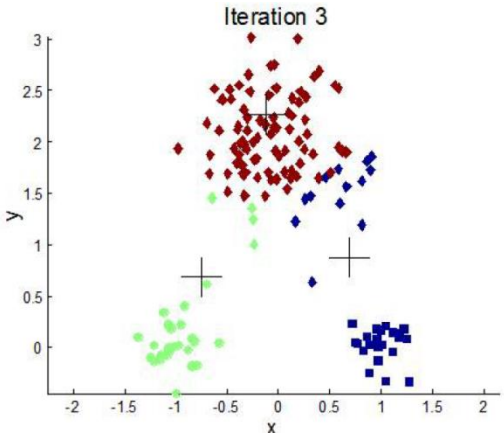
De esta forma, el algoritmo de las k-medias :

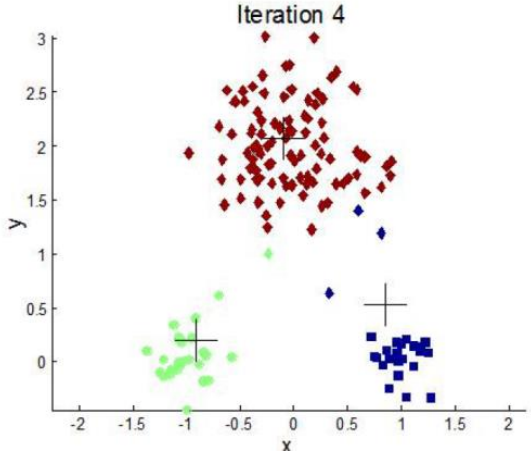
1. Elegirá en primer lugar un conjunto de puntos denominados centroides, tantos como clusters k se hayan especificado al inicializar el algoritmo, de manera que cada uno de esos centroides representa un elemento prototípico de cada uno de los k grupos.
2. Posteriormente, se calculará la distancia desde cada punto a cada uno de los centroides, de manera que se asignará cada punto al centroide que minimiza su distancia.
3. Una vez se han construido los nuevos grupos, se calcularán los nuevos centroides atendiendo a algún criterio, como puede ser la media de los elementos que pertenecen al grupo.

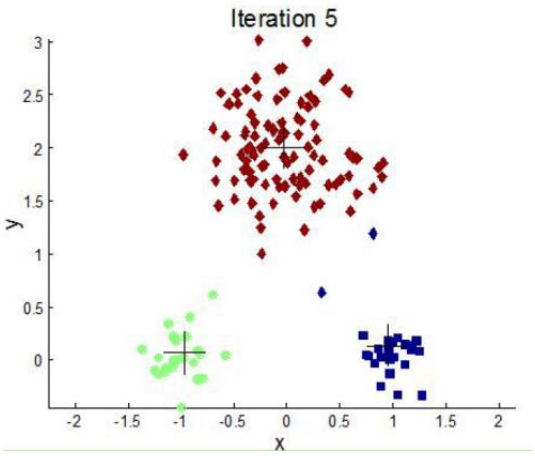
Este funcionamiento puede verse de manera gráfica en la siguiente secuencia de imágenes, que resumen el procedimiento de actuación del algoritmo.

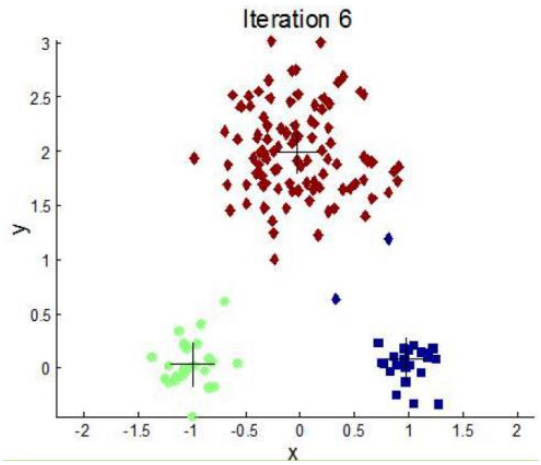












En cada una de las sucesivas iteraciones, los elementos van re-asignándose a un grupo u otro en función de la posición de los centroides, que se actualizan también en cada iteración.

De manera más formal a las representaciones en pseudo-código vistas anteriormente, el algoritmo de las k-medias puede formalizarse de la siguiente forma:

Algoritmo de las K-medias (K-means)

1.Sean $\{x_1, x_2, \dots, x_n\}$ n ítems definidos en un espacio d-dimensional E, con una matriz de datos $x_{il}, i = \{1, \dots, n\}$ y $l = \{1, \dots, d\}$ y distancia $p(.,.)$. Elegir K y un conjunto $c_1, \dots, c_k$ de centroides iniciales. Sea $\{G_1, \dots, G_k\}$ el conjunto de grupos que vamos a obtener, inicialmente $G_j = \emptyset \; \forall j \in \{1, .., K\}$.

2. $\forall i \in \{1, \dots, n\}$:

   2.1. Calcular $j_i \in \{1, \dots, K\}; p(x_i, c_{ji}) = \min_{j \in \{1,..,k\}} = (p(x_i, c_j))$

   2.2 Hacer $G_{ji} = G_{ji} \cup \{x_i\}$

3. Obtener los nuevos centroides haciendo:

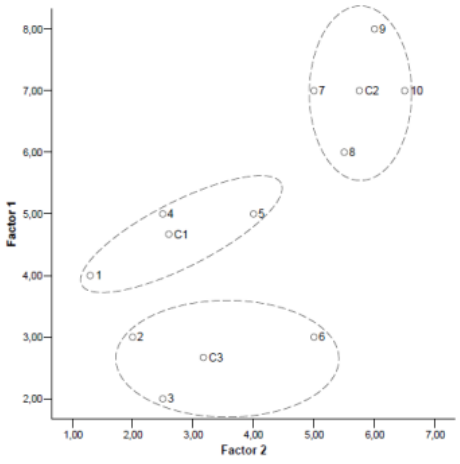
   $\forall j \in \{1, \dots, K\}, \; \forall j \in \{1, \dots, d\} cn_{jl} = \sum_{x_i \in G_j} \frac{x_{il}}{|G_j|}$

4.Si $cn_j = c_j \forall j \in \{1, \dots, k\}$, parar. En caso contrario:

   4.1. Hacer $cn_j = c_j \forall j \in \{1, \dots, K\}$

   4.2. Hacer $G_j = \emptyset \; \forall j \in \{1, \dots, K\}$

Los resultados de este algoritmo serán finalmente la distribución de los diferentes ítems del dataset en los k grupos determinados a priori en la ejecución del k-nn. Un ejemplo del resultado del algoritmo de las k-medias sobre un conjunto de diez elementos, es el que se muestra a continuación.



Para concluir con el estudio del algoritmo de las k-medias, es importante citar las principales ventajas que proporciona este método, a la par que valorar sus inconvenientes o debilidades, con el objetivo de saber si el algoritmo es adecuado o no para cierto tipo de problemas, de cara al momento en que se nos plantee un problema en el futuro. La siguiente tabla analiza cada uno de estos factores.

Algoritmo de las K-medias	
Fortalezas	**Debilidades**
Computacionalmente eficiente **O(tkn)** t: Nº de Iteraciones k: Nº de Clusters n: Nº de elementos	Necesidad de Especificar K
Normalmente, $k \wedge t \ll n$	Problemas de Inicialización
Fácil de Implementar	Pueden aparecer grupos vacíos
	Diferentes formas, tamaños,...

Tal y como se puede observar en la tabla, el algoritmo de las k-medias es fuertemente dependiente de los parámetros de entrada, es decir, del número de clusters seleccionados así como de los centroides inicializados, siendo la inicialización de estos parámetros uno de los problemas más importantes de este algoritmo y uno de los más estudiados

Por este motivo una buena medida de bondad del agrupamiento, es decir, que mida cómo de bueno ha sido el agrupamiento realizado por el algoritmo, consiste en la suma total de la proximidad que se minimiza.

$$SSE = \sum_{j=1}^{j=K} \sum_{x_i \in G_j} p(x_i, c_j)/n$$

En caso de trabajar con la distancia euclídea, de ésta fórmula se obtiene el error cuadrático global, el cual será menor cuantos más grupos se consideren, aunque puede ser que esta opción no sea la más adecuada. Por este motivo, **uno de los procedimientos más utilizados consiste en realizar en primer lugar un agrupamiento jerárquico que aporte una idea preliminar para justar estos parámetros.**

Un ejemplo en el que se utiliza la ayuda que proporciona el enfoque jerárquico es el que aparece en las transparencias y que se cita a continuación.

EJEMPLO: Agrupamiento en tres grupos proporcionado por el enfoque jerárquico usando el método de enlace completo.

$$P_1 = \{\{1,2,3\},\{4,5,6\},\{7,8,9,10\}$$

Agrupamiento en cinco grupos, resultado de aplicar el método de enlace simple:

$$P_2 = \{\{1,2,3\},\{4\},\{5\},\{6\},\{7,8,9,10\}\}$$

La siguiente tabla muestra los valores de SSE para cada caso:

N. de grupos	Sin selección previa	Con selección previa
3	1.083	0.980
5	0.650	0.590

A la luz de los resultados de la tabla se puede concluir:

1. **El valor del SSE disminuye conforme aumenta el número de grupos, aunque esta solución puede no ser la mejor, tal y como se dijo anteriormente.**
2. **La elección inicial utilizando un agrupamiento obtenido por un algoritmo jerárquico mejora esta medida de bondad.**

4.2. DBSCAN

Tal y como y se vio en la introducción, el algoritmo de agrupamiento **DBSCAN** es un algoritmo de Clustering Particional basado en **densidad**, que trata de buscar zonas de alta densidad rodeadas por zonas de baja densidad. Entre las características generales y más importantes de este algoritmo podemos destacar:

1. Sigue un enfoque de **densidad basada en centros**.
2. La densidad se estima para un punto concreto **contando el número de puntos que caen dentro de un entorno centrada en el mismo y de un radio fijado (eps).**
3. La densidad de cada punto **depende del radio que se tome**, si el radio es suficientemente grande, la densidad de cada punto es n; por el contrario si es suficientemente pequeño la densidad es de 1 en cada punto.

Una vez identificadas las características generales del algoritmo, estamos en disposición de describir su funcionamiento.

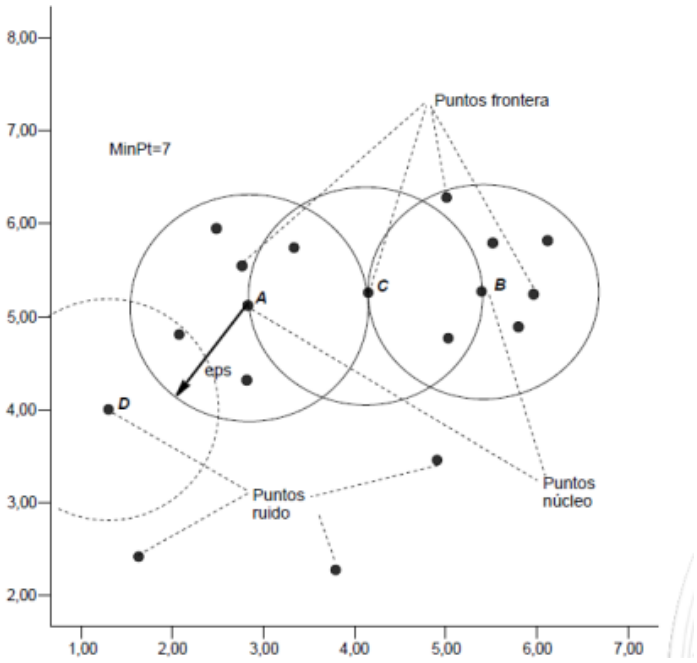
Funcionamiento DBSCAN

1. En primer lugar, se colocan en un mismo grupo todos los puntos núcleo que distan entre sí menos de eps, utilizando un criterio de transitividad:
   a. Si la distancia entre dos puntos núcleo $f(n_1, n_2) \leq eps$ y para otro punto núcleo $f(n_2, n_3) \leq eps$ entonces $n_1, n_2 y\ n_3$ pertenecen al mismo grupo.
   b. Se dice entonces que $n_1 y\ n_2$ o $n_2 y\ n_3$ son directamente densidad alcanzables mientras que $n_1 y\ n_3$ se dice que son densidad alcanzables.
2. También se asignarán al mismo grupo todos los puntos frontera asociados a cada punto núcleo.
3. Se eliminan los puntos de ruido.

En realidad, **DBSCAN** funciona de manera **iterativa**, y una vez fijados los parámetros **eps** y **MinPT**:

1. Va considerando punto a punto, estableciendo su entorno de radio eps y viendo si es o no un núcleo, construyendo un grupo con dicho entorno en caso de que lo sea.
2. Se buscan otros núcleos que sean densidad alcanzables a partir de él, si existe alguno. En caso de que exista alguno, el grupo generado inicialmente se une a aquel al que pertenezca este núcleo.
3. El proceso termina cuando ningún punto puede ser añadido a ningún grupo.

De manera gráfica, la siguiente imagen proporciona una pequeña traza del algoritmo DBSCAN:



Julio Alberto López Gómez
Amparo Vila Miranda

De manera más formal, el siguiente fragmento de código especifica el pseudocódigo del algoritmo DBSCAN:

Algoritmo DBSCAN

1.Fijar

1.1. Un valor de radio **eps**

1.2. Un número de puntos mínimo: **MinPt** adecuado para que se considere que entorno tiene densidad suficiente para formar parte de un grupo

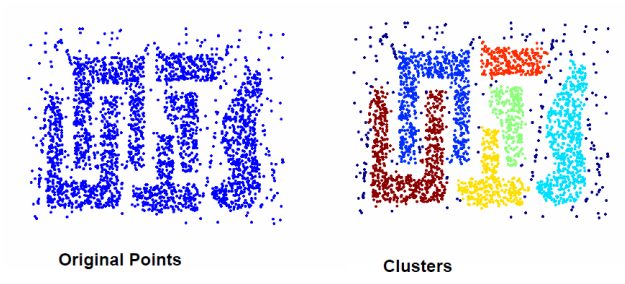
2. Todo punto del espacio de patrones se puede clasificar como:

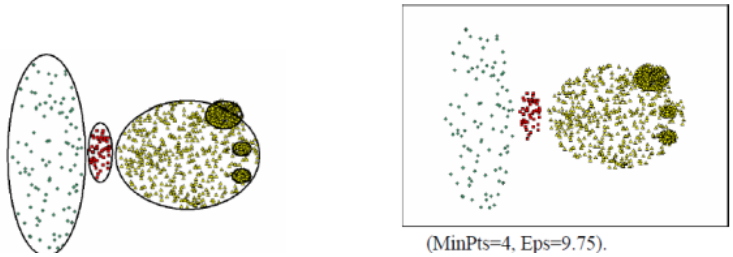
2.1. Punto Núcleo: Son aquellos puntos que se considera que pertenecen al interior de un grupo y se definen como aquellos que son centro de un entorno de eps que tiene más de MinPt puntos.

2.2. Punto Frontera: Son aquellos que se encuentran en un entorno de radio eps que tiene como centro un punto núcleo, puede ocurrir que un punto frontera pertenezca al entorno de varios puntos núcleo.

2.3. Punto Ruido: Son aquellos puntos que no son núcleo ni frontera. Se supone que va a estar en regiones muy poco densas y que no va a formar parte de ningún grupo.

A continuación, se muestran dos ejemplos en los que se muestran, respectivamente, una ejecución correcta de DBSCAN y una ejecución en la que el algoritmo no da unos buenos resultados.





Por último, y para analizar la eficiencia del algoritmo, la siguiente tabla muestra una comparativa de las principales fortalezas y debilidades de DBSCAN, con el objetivo de poder decidir si es o no adecuado aplicar este tipo de algoritmos a un problema dado.

Algoritmo DBSCAN	
Fortalezas	**Debilidades**
Resistente al ruido	No puede manejar densidades variables
Maneja clusters de diferentes formas y tamaños	Sensible a parámetros complejos de determinar

No obstante, existen generalizaciones de DBSCAN que permiten evitar en lo posible estos problemas. **OPTICS** debido a Ankerst et. Al trabaja con regiones de densidad variable, considerando valores de eps menores.

En conclusión, DBSCAN es un algoritmo potente y sencillo que puede optimizarse para espacios e baja dimensionalidad. Produce grupos complejos para los cuales no hay ninguna hipótesis de "centralidad" o "globularidad" como en el caso del método de las k-medias.

5. Validación de Clusters

El problema de la validación de clusters ha sido ampliamente analizado y estudiado por los investigadores del campo de la Inteligencia Artificial. El problema que ocupa a la validación de clusters es:

¿Cómo puedo evaluar la bondad de los resultados de mi algoritmo de agrupamiento?

O lo que es lo mismo, ¿Cómo puedo saber cuán bueno es el agrupamiento que he realizado?

Es necesario recordar en este punto, que en principio, no se conoce nada acerca del problema, pues nos hayamos en contexto no supervisado. La evaluación de los clusters servirá para:

- **No confundir grupos con ruido.**
- **Comparar los resultados de diferentes algoritmos de agrupamiento.**
- **Comparar dos conjuntos de clusters .**
- **Comparar dos clusters.**

Además, la validación cluster o validación de agrupamientos será útil para:

1. Determinar la tendencia de agrupamiento de un conjunto de datos, es decir, distinguir si la estructura no aleatoria realmente existe en los datos.
2. Comparar los resultados de un análisis cluster con una partición conocida previamente.
3. Evaluar cómo los resultados de un análisis cluster se ajustan a los datos sin referencia a información externa.
4. Comparar los resultados de dos conjuntos diferentes de análisis cluster para decidir cuál es mejor.
5. Determinar el número de grupos correcto.

Medidas para evaluar Agrupamientos.

Las medidas para evaluar agrupamientos son medidas que se usan para determinar distintos aspectos de la validez de un cluster.

Estas medidas son de diversos tipos:

- **Medidas no Supervisadas**: Se utilizan para medir la bondad de un agrupamiento sin tener ninguna información adicional al respecto. El ejemplo más clásico es el **error cuadrático** que se aplica como medida de bondad en el algoritmo de las k-medias. A su vez, las medidas no supervisadas pueden dividirse en:
  - **Medidas de Cohesión:** Miden cómo de compactos son los grupos.
  - **Medidas de Separación:** Miden cómo están de separados los grupos.

  Estas medidas se denominan también índices Internos ya que sólo utilizan los datos del problema.

  Las medidas de cohesión y separación se calculan de forma diferente según se hayan obtenido los grupos mediante técnicas basadas en prototipos o se hayan usado otras técnicas.

  - Cuando **no** se dispone de **prototipo** (centroide) se tiene:

$$cohesion(C_i) = \sum_{x,y \in C_i} proximity(x,y)$$

$$separacion(C_i, C_j) = \sum_{x \in C_i, y \in C_j} proximity(x,y)$$

  - Cuando se dispone un **centroide** c_i de cada grupo, entonces:

$$cohesion(C_i) = \sum_{x \in C_i} proximity(x, c_i)$$

$$separacion(C_i, C_j) = proximity(c_i,)$$

  - **Coeficiente de Silueta (Silhouette Coefficient)**: Medida que combina las ideas de separación y cohesión. Se calcula para cada punto de la siguiente forma:
    - Se calcula la **distancia media del punto i a los elementos de su grupo (a_i)**
    - Se calcula la **media de la distancia de i a los elementos de cada grupo que no es el suyo y se hace el mínimo de todas estas medias. Nótese como b_i.**
    - El coeficiente de Silueta para i viene dado por:

$$s_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}$$

- Este coeficiente varía entre -1 y 1. **Es deseable un coeficiente positivo y cerca de 1**.
- **Para calcular el coeficiente de silueta de un grupo o un agrupamiento se calcula mediante la media de los correspondientes coeficientes .**

- **Medidas Supervisadas:** Medidas de validación que miden la adecuación del agrupamiento obtenido como una partición ya existente. Las medidas supervisadas se denominan también Índices externos ya que utilizan información no presente en el data set. En la validación supervisada, existen dos enfoques principalmente.
  - **Orientado a la clasificación:** Miden cómo se ajusta la partición obtenida en un agrupamiento a una clasificación previamente dada.
    Sean $\{G_i; i \in \{1, \dots, K\}\}$ los grupos obtenidos y $\{C_j; j \in \{1, \dots, L\}\}$ las clases a comparar, sea m el número total de puntos. Se define:
    $$\forall i \in \{\{1, \dots, K\}; j \in \{1, \dots, L\}\} p_{ij} = \frac{m_{ij}}{m_i}$$
    Donde m_{ij} es el número de ítems que hay de la clase C_j en el grupo G_i y m_i es el número de ítems que hay en el grupo G_i. p_{ij} es la probabilidad de que un miembro de G_i pertenezca a C_j.
    A partir de p_{ij} se definen:
    - **Entropía:** $\forall i \in \{1, \dots, K\} e_i = -\sum_{j=1}^{L} p_{ij} \log_2 p_{ij}$ es la entropía de un cluster $e = \frac{\sum_{i=1}^{K} m_i e_i}{m}$ es la entropía total.
    - **Purity:** La pureza de un grupo es $p_i = \max_j(p_{ij})$ la total de agrupamiento es $purity = \frac{\sum_{i=1}^{K} m_i p_i}{m}$
    - **Precisión y Recall:** $precision(i,j) = p_{ij}; recall(i,j) = \frac{m_{ij}}{n_j}$ donde n_j es el número de elementos de la clase C_j
    - **F-medida:** $F(i,j) = \frac{2xprecision(i,j)xrecall(i,j)}{precision(i,j)+recall(i,j)}$

  - **Orientado a la similaridad:** La idea básica de este enfoque es construir matrices de incidencia del agrupamiento:
    $$\text{IG}_{ij} \begin{cases} 1 \text{ si i j están en el mismo cluster} \\ 0 \text{ en caso contrario} \end{cases}$$
    y de la clasificación:
    $$IC_{ij} = \begin{cases} 1 \; si \; i \; j \; están \; en \; la \; misma \; clase \\ 0 \; en \; caso \; contrario \end{cases}$$

    Y establecer medidas de coincidencia entre ambas. Una de las más comunes es calcular la **correlación** entre ambas matrices.
    Otra alternativa consiste en crear la denominada matriz de confusión, según la cantidad de parejas de puntos que coincide o difieren, tal y como se muestra en la siguiente tabla.

	Igual cluster	Diferente cluster
Igual clase	f_{11}	f_{10}
Diferente clase	f_{01}	f_{00}

En base a la matriz de confusión anterior, se definen los siguientes coeficientes o estadísticos:

- **Coeficiente de Jaccard:** $\frac{f_{11}}{f_{10}+f_{01}+f_{00}}$
- **Estadístico de Rand**: $\frac{f_{11}+f_{00}}{f_{10}+f_{01}+f_{00}+f_{01}}$

- **Medidas Relativas:** Se utilizan para comparar diferentes agrupamientos o grupos. Pueden ser supervisadas o no pero siempre se formularán de forma relativa con el objetivo de comparación. Por ejemplo, el uso de la medida SSE para comparar el proceso de selección de grupos iniciales que se vio en el método de las k-medias.

En general, se considera que la medida de un agrupamiento es la suma ponderada de la medida de sus grupos. Es decir, la validez total de un agrupamiento de K grupos es:

$$\boldsymbol{Validez\ Total} = \sum_{i=1}^{K} \boldsymbol{w_i validez(C_i)}$$

Donde la validez puede ser una medida de cohesión, separación o una combinación de ambos y los pesos pueden ser 1, el tamaño del cluster o expresiones más sofisticadas.

6. Extensiones de los métodos anteriores

Los métodos anteriores presentan diferentes problemas, lo cual ha hecho necesaria la aparición de diversos métodos que extienden a los anteriores para solucionar dichos problemas.

6.1. Extensiones de los métodos jerárquicos

PROBLEMA: El problema principal, se encuentra en que las técnicas de agrupamiento jerárquico siempre han sido muy costosas desde el punto de vista computacional (Complejidad mínimo $O(n^2)$). Este inconveniente se agrava ya que, en su versión clásica, estas técnicas obtienen el conjunto de todos los posibles agrupamientos, desde el inicial con n grupos hasta el último con un solo grupo (Complejidad en el peor caso $O(2^n)$).

Las priemras versiones que trataron de evitar estos inconvenientes fueron los algoritmos llamados **DIANA** Y **AGNES** de Kauffman y Rosseeuw respectivamente, surgidos en 1990. El primero de ellos presenta un enfoque divisivo y el segundo aglomerativo, habiendo de especificar en ambos el número de grupos que se desea como una condición de terminación.

PROBLEMA: Otro problema de los algoritmos tradiciones es el hecho de que, cuando se toma la decisión de unir dos grupos en los enfoques aglomerativos (equivalente a dividir un grupo en los algoritmos divisivos), no se puede volver atrás.

Este sentido, la solución consiste en mejorar la calidad de los grupos obtenidos utilizando otras técnicas de agrupamiento, realizando un proceso de múltiples fases. Entre las soluciones más populares se encuentran:

- **BIRCH (Balanced Iterative Reducing and Clustering using Hierchies):** Zhang et. Al 1966. Este método inicialmente particiona los patrones de forma jerárquica utilizando estructuras de **árboles** y posteriormente aplica otros algoritmos de agrupamiento para refinar el resultado. Este algoritmo es **fácilmente escalable** y es muy utilizado en problemas de Minería de Datos. Su funcionamiento se basa en el uso de **resúmenes de los datos** (estadísticos suficientes) para sustituir a los datos originales.
- **CURE (Clustering Using Representatives):** Guha et. al. 1998. Este método, por su parte, emplea **agrupamiento jerárquico** que se encuentra a mitad de camino entre los métodos de enlace ponderado y de enlace completo, ya que en lugar de utilizar un único punto, o todos ellos para representar un grupo, **elige un número fijo de puntos representativos**. Estos puntos se generan eligiendo primeramente un conjunto de puntos bien distribuidos sobre el grupo y posteriormente **reduciendo la distancia de estos al centro del grupo un determinado factor llamado factor de reducción**. Los grupos con puntos representativos más cercanos se unen en cada paso del algoritmo.
- **ROCK:** Se trata de un algoritmo aglomerativo desarrollado por los mismos autores que el anterior, orientado al uso de atributos categóricos.
- **CHAMELEON:** Karpys et. Al. 1999. Algoritmo que explora un modelo dinámico de agrupamiento jerárquico. En su proceso de agrupamiento, dos grupos se unen si la interconectividad y la cercanía entre ellos se corresponde con la conectividad interna y la cercanía de los ítems que están en los grupos.

6.2. El algoritmo de las k-medias difuso

Al menos implícitamente, en todos los métodos de agrupamiento clásicos **se ha supuesto la hipótesis de que el agrupamiento es exclusivo, es decir, los patrones se particionan en conjuntos disjuntos**. Esta suele ser la mejor opción cuando los grupos son compactos y están bien separados. No obstante, **el problema aparece cuando los grupos tienen puntos comunes o incluso se solapan. Los conjuntos de este tipo son conjuntos cuyas fronteras están mal definidas o "borrosas".**

La teoría de subconjuntos difusos (fuzzy sets), permite que un patrón pertenezca a un grupo con un cierto "grado de pertenencia".

Conceptos Básicos sobre Conjuntos difusos.

- Cada grupo o conjunto difuso $C_j: j \in \{1, \dots, K\}$ tiene asociada una "función de pertenencia".

$$C_j : X \rightarrow [0,1]$$

  Siendo $X = \{x_1, x_2, \dots, x_n\}$ el espacio de patrones.

- El valor $u_{ij} = C_j(x_i)$ mide el grado de pertenencia del punto x_i al grupo C_j. Los valores u_{ij} constituyen la "matriz de pertenencia" que notaremos U.
- Es habitual imponer la condición de partición difusa o posibilística:

$$\sum_{j=1}^{K} u_{ij} = 1, \forall i \in \{1, \dots, N\}, \max_{j \in \{1,..,K\}} u_{ij} = 1, \forall i \in \{1, \dots, N\}$$

Una vez asimilados estos conceptos fundamentales acerca de la teoría de conjuntos difusos, es posible señalar una serie de características de este algoritmo.

- **El grado de pertenencia no tiene el mismo significado que una probabilidad**, ya que el grado de pertenencia se puede interpretar como una medida de compatibilidad entre el punto x_i con el grupo C_j, entendido este como el resultado de una propiedad o conjunto de propiedades expresadas de forma imprecisa.
- Este enfoque es **muy útil cuando se intenta hacer una interpretación de los grupos**, ya que, en muchos casos, las descripciones de los grupos obtenidos en un problema concreto serán de tipo impreciso por selo las etiquetas que los caracterizan.
  - **EJEMPLO:** Agrupar un conjunto de coche intentando obtener aquellos de gama alta, media o utilitario
  - **EJEMPLO2:** Agrupar un conjunto de finca o parcelas atendiendo a los cultivos que se plantan sobre ellas.
- La gran mayoría de los algoritmos de agrupamiento basados en conjuntos difusos hacen uso del concepto de **partición difusa.**
- Por otra parte, **existen distintos enfoques para diseñar algoritmos particionales de tipo difuso, la mayor parte de los cuales son generalizaciones más o menos directas del método de las k-medias**.

Una vez definidos completamente los conceptos básicos sobre lógica difusa necesarios para construir el algoritmo, así como sus características principales, estamos en disposición de describir el algoritmo utilizando una representación formal cercana al pseudocódigo.

Algoritmo de las K-Medias Difuso

1. Seleccionar una partición difusa inicial de N objetos en K grupos seleccionando una matriz de pertenencia U.

2. Calcular los "centros" de los grupos difusos asociados a U mediante la expresión: $c_j = \sum_{i=1}^{N} u_{ij} x_i$

3. Calcular el valor óptimo de:

$$E^2(U) = \sum_{i=1}^{N} \sum_{j=1}^{K} u_{ij} ||x_i - c_k||^2$$

De esta forma se obtiene un problema de optimización sobre los valores de pertenencia u_{ij}, sujetos a:

$$\sum_{j=1}^{K} u_{ij} = 1; u_{ij} \geq 0 \text{ o bien } \max_{j \in \{1,\dots,K\}} u_{ij} = 1; u_{ij} \geq 0$$

4. Repetir desde el paso 2 hasta que los valores de U no cambien significativamente.

El algoritmo mostrado en el fragmento anterior, es una generalización del método de las K-medias difuso. No obstante, cabe destacar que existen una serie de variantes sobre este algoritmo bastante extendidas y que se analizan a continuación.

VARIANTES. Algoritmo de las K-Medias Difuso.

1. **El centro de un grupo difuso no es su media sino su valor más representativo, el cual viene dado por:**

$$\forall j \in \{1, \dots, K\}; c_j = x_{lj} | u_{lj} = \max_{i \in \{1,\dots,n\}} u_{ij}$$

2. Utilización de **otras funciones de distancia** más generales.
   a. La distancia entre dos puntos asociada al grupo C_j podría expresarse:

$$\forall x, y, d_j(x, y) = \min\left(u_j(x), u_j(y)\right) ||y - x||^2$$

   b. De esta forma, la distancia se transforma en:

$$S^2(U) = \sum_{i=1}^{N} \sum_{j=1}^{K} d_j(x_i, c_j)$$

   c. Así, si $u_{ij}(c_j) = 1$ obtenemos la expresión inicial.

3. Existen otras variantes que utilizan una **función de distancia adaptativa**.

Julio Alberto López Gómez
Amparo Vila Miranda

6.3. Los métodos de k-medioides

El método de los k-medioides, surge debido a que en el método de las k-medias, se supone que el espacio de ítems es continuo y por tanto, que el centroides puede ser un posible ítem. Sin embargo, esto solo sucede cuando todos los datos son numéricos y continuos.

Una alternativa al uso del centroide que toma como prototipo de cada grupo un punto del mismo que se considera representativo del mismo, es la utilización del medioide, como punto del grupo que se considera representativo. Todo el proceso iterativo del método de las k-medias se realiza minimizando las sumas de las distancias de cada punto a los medioides considerados.

Los principales métodos de k-medioides desarrollados y más utilizados son:

- **PAM: Kuffmann y Rousseau 1990.** En este método se parte de una selección inicial de k medioides. Dichos medioides generarán una partición en k grupos del conjunto total de ítems. Después, para cada grupo obtenido se intenta reemplazar el medioide asociado a él por algún punto del mismo grupo que sea más idóneo. Esto se hace considerando cada punto del grupo y calculando la distancia total del resto de los elementos del grupo a dicho punto. En caso de que ésta mejore la del medioide, este es sustituido por el punto en cuestión. Finalmente, los grupos son recalculados en base a los nuevos medioides seleccionados, repitiendo sucesivamente este procedimiento.
- **CLARA: Kauffmann y Rousseau 1990.** El principal problema del algoritmo PAM es que no se adapta bien a grandes bases de datos. Para solucionar este problema nació CLARA, algoritmo basado en un proceso de muestreo. CLARA realiza sucesivos muestreos aplicando PAM a cada uno de ellos. Los conjuntos de medioides obtenidos se aplican al conjunto total, seleccionando aquel que nos dé una menor distancia global. Con estos puntos de partida se genera un nuevo proceso de muestreo y una nueva iteración. La complejidad de cada iteración es de $O(kS^2 + k(n-k))$, donde S es el tamaño de la muestra. La efectividad de CLARA depende del tamaño de la muestra, si bien es fácilmente escalable-
- **CLARANS (NG y Han 1994).** En este método se considera la búsqueda de los medioides óptimos como un proceso de búsqueda en un árbol donde cada nodo es un conjunto de k- medioides. Así, considerando un nodo, el agrupamiento obtenido reemplazando un medioide del mismo por algún otro punto se denomina entorno. El proceso de este algoritmo prueba una serie de entornos generando puntos aleatoriamente, si encuentra un entorno mejor que el agrupamiento considerado, el algoritmo se mueve a este nodo y comienza de nuevo a probar, en caso contrario se considera que se ha llegado a un óptimo local.
  Cuando el algoritmo llega a un óptimo local, éste comienza con un nuevo conjunto de nodos obtenidos por medio de un muestreo aleatorio y una aplicación de PAM. El algoritmo termina cuando se ha alcanzado un número suficiente de mínimos locales (generalmente 2). La complejidad de Clarans es de $O(n^2)$. Posteriormente, Ester,Kriegel y XU 1995 mejoraron este algoritmo mediante el uso de árboles R*.

Bibliografía y Referencias.

[1] Pang- Ning Tan. Introduction to Data Mining.

[2] A. Vila. Introducción a las técnicas de agrupamiento (Clustering). Material de clase.

[3] J. Gao. Clustering Lecture 3: Hierarchical Methods. Transparencias.

[4] K. Chen. Hierarchical Clustering. Transparencias.

[5] Ian H. Witten Eibe Frandk. Data Mining Practical Mahine Learning Tools and Techniques, Second Edition Elsevier (2005).