Actividades resueltas

Algoritmo de *clustering* jerárquico

Descripción de la actividad

La siguiente base de datos contiene cinco registros sobre la composición de diferentes tipos de vidrios. Para cada vidrio se indica el porcentaje en el peso de cada uno de los elementos químicos que lo componen: Na, Si y Ca.

	Na	Si	Ca
A	12.8	73	8.8
B	12.2	72.9	8.6
C	12.8	73.3	8.8
D	13.6	73	8.9
E	13.1	72.9	9.1

A partir de esta base de datos se quiere agrupar los diferentes vidrios de forma que se obtengan clústeres de vidrios similares. Para ello se va a aplicar el algoritmo de **clustering jerárquico aglomerativo** utilizando la medida de distancia de **enlace sencillo**.

El primer paso es calcular la matriz de distancias entre las diferentes instancias. Se ha computado ya parte de esta matriz:

	A	B	C	D	E
A	0	0.64	0.30	0.81	0.44
B	0.64	0	0.75	1.44	1.03
C	0.30	0.75	0		
D	0.81	1.44		0	
E	0.44	1.03			0

Acaba de calcular la matriz de distancias y aplica el algoritmo de clustering jerárquico aglomerativo. Describe claramente los pasos que se realizan en la ejecución del algoritmo. Además representa la estructura jerárquica de clústeres obtenida en un dendograma.

Resolución de la actividad

Tal como se indica en el enunciado primero debemos acabar de calcular la **matriz de distancias**. Para ello se calcula la distancia Euclídea entre la instancia C y la instancia D, entre la instancia C y la instancia E, y entre la instancia D y la instancia E.

Cada instancia se modela como un vector de tres dimensiones con los valores de los tres atributos que componen el registro. Por tanto una instancia I_i de la base de datos de ejemplos tendrá la forma (Na, Si, Ca). Entonces la instancia C se representa con el vector (12.8, 73.3, 8.8), la instancia D con el vector (13.6, 73, 8.9) y la instancia E con el vector (13.1, 72.9, 9.1).

La fórmula para calcular la distancia Euclídea entre dos instancias I_i y I_j es la siguiente:

$$d(I_i, I_j) = \sqrt{(Na_i - Na_j)^2 + (Si_i - Si_j)^2 + (Ca_i - Ca_j)^2}$$

Por tanto, la distancia entre la instancia C y la instancia D es:

$$d(C, D) = \sqrt{(12.8 - 13.6)^2 + (73.3 - 73)^2 + (8.8 - 8.9)^2} = \sqrt{0.8^2 + 0.3^2 + 0.1^2} = 0.86$$

La distancia entre la instancia C y la instancia E es:

$$d(C, E) = \sqrt{(12.8 - 13.1)^2 + (73.3 - 72.9)^2 + (8.8 - 9.1)^2} = \sqrt{0.3^2 + 0.4^2 + 0.3^2} = 0.58$$

La distancia entre la instancia D y la instancia E es:

$$d(D, E) = \sqrt{(13.6 - 13.1)^2 + (73 - 72.9)^2 + (8.9 - 9.1)^2} = \sqrt{0.5^2 + 0.1^2 + 0.2^2} = 0.55$$

Una vez se han calculado estas tres distancias Euclídeas se puede completar la matriz de distancias con estos valores. Además teniendo en cuenta la matriz es simétrica se obtiene el siguiente resultado:

	A	B	C	D	E
A	0	0.64	0.30	0.81	0.44
B	0.64	0	0.75	1.44	1.03
C	0.30	0.75	0	0.86	0.58
D	0.81	1.44	0.86	0	0.55
E	0.44	1.03	0.58	0.55	0

A partir de esta matriz de distancias se empieza a aplicar el algoritmo de clustering jerárquico aglomerativo, generando iterativamente los clústeres para cada nivel.

En el **Nivel 0** se crea un clúster para cada una de instancias. Entonces tenemos 5 clústeres:

$$C_A = \{A\}$$
$$C_B = \{B\}$$
$$C_C = \{C\}$$
$$C_D = \{D\}$$
$$C_E = \{E\}$$

En la **primera iteración** se busca el par de clústeres menos distantes. Comprobando la matriz de distancias se observa que la menor distancia se encuentra entre A y C. Estas dos instancias se encuentran a distancia 0.30.

Por lo tanto se unen los clústeres $C_A = \{A\}$ y $C_C = \{C\}$ en el nuevo clúster $C_{AC} = \{A, C\}$.

Finalmente se recalcula la matriz de distancias. Para ello se eliminan las filas y columnas correspondientes a C_A y C_C, se añade una nueva fila y una nueva columna para C_{AC} y se recalculan las distancias mediante la medida de enlace sencillo. Como se está utilizando enlace sencillo la distancia entre el clúster C_{AC} y el clúster C_i será el mínimo de la distancia entre el clúster C_i y el clúster C_A y la distancia entre el clúster C_i y el clúster C_C.

	$C_A = \{A\}$	$C_B = \{B\}$	$C_C = \{C\}$	$C_D = \{D\}$	$C_E = \{E\}$	$C_{AC} = \{A,C\}$
$C_A = \{A\}$	0	0.64	0.30	0.81	0.44	
$C_B = \{B\}$	0.64	0	0.75	1.44	1.03	min{0.64, 0.75} =0.64
$C_C = \{C\}$	0.30	0.75	0	0.86	0.58	
$C_D = \{D\}$	0.81	1.44	0.86	0	0.55	min{0.81, 0.86} =0.81
$C_E = \{E\}$	0.44	1.03	0.58	0.55	0	min{0.44, 0.58} =0.44
$C_{AC} = \{A,C\}$		-		-	-	0

En base a estos cálculos, la matriz de distancias obtenida sería la siguiente:

	$C_B = \{B\}$	$C_D = \{D\}$	$C_E = \{E\}$	$C_{AC} = \{A,C\}$
$C_B = \{B\}$	0	1.44	1.03	0.64
$C_D = \{D\}$	1.44	0	0.55	0.81
$C_E = \{E\}$	1.03	0.55	0	0.44
$C_{AC} = \{A,C\}$	0.64	0.81	0.44	0

Por tanto en el **Nivel 1** tendremos cuatro clústeres:

$$C_B = \{B\}$$
$$C_D = \{D\}$$
$$C_E = \{E\}$$
$$C_{AC} = \{A, C\}$$

Como no todas las instancias forman parte del mismo clúster se repite el proceso.

Entonces en la **segunda iteración** se busca el par de clústeres menos distantes que serían el clúster C_E y el clúster C_{AC} que se encuentran a distancia 0.44.

Por lo tanto se unen los clústeres $C_E = \{E\}$ y $C_{AC} = \{A, C\}$ en el nuevo clúster $C_{ACE} = \{A, C, E\}$.

Finalmente se recalcula la matriz de distancias. Para ello se eliminan las filas y columnas correspondientes a C_E y C_{AC}, se añade una nueva fila y una nueva columna para C_{ACE} y se recalculan las distancias mediante la medida de enlace sencillo.

	$C_B = \{B\}$	$C_D = \{D\}$	$C_E = \{E\}$	$C_{AC} = \{A,C\}$	$C_{ACE} = \{A,C,E\}$
$C_B = \{B\}$	0	1.44	1.03	0.64	0.64
$C_D = \{D\}$	1.44	0	0.55	0.81	0.55
$C_E = \{E\}$	1.03	0.55	0	0.44	
$C_{AC} = \{A,C\}$	0.64	0.81	0.44	0	
$C_{ACE} = \{A,C,E\}$	-	-			0

En base a estos cálculos, la matriz de distancias obtenida sería la siguiente:

	$C_B = \{B\}$	$C_D = \{D\}$	$C_{ACE} = \{A,C,E\}$
$C_B = \{B\}$	0	1.44	0.64
$C_D = \{D\}$	1.44	0	0.55
$C_{ACE} = \{A,C,E\}$	0.64	0.55	0

Por tanto en el **Nivel 2** tendremos tres clústeres:

$$C_B = \{B\}$$

$$C_D = \{D\}$$

$$C_{ACE} = \{A, C, E\}$$

Como no todas las instancias forman parte del mismo clúster se repite el proceso.

En la **tercera iteración** se busca el par de clústeres menos distantes que serían el clúster C_D y el clúster C_{ACE} que se encuentran a distancia 0.55.

Por lo tanto se unen los clústeres $C_D = \{D\}$ y $C_{ACE} = \{A, C, E\}$ en el nuevo clúster $C_{ACED} = \{A, C, E, D\}$.

Finalmente se recalcula la matriz de distancias. Para ello se eliminan las filas y columnas correspondientes a C_D y C_{ACE}, se añade una nueva fila y una nueva columna para C_{ACED} y se recalculan las distancias mediante la medida de enlace sencillo.

	$C_B = \{B\}$	$C_D = \{D\}$	$C_{ACE} = \{A,C,E\}$	$C_{ACED} = \{A,C,E,D\}$
$C_B = \{B\}$	0	1.44	0.64	0.64
$C_D = \{D\}$	1.44	0	0.55	
$C_{ACE} = \{A,C,E\}$	0.64	0.55	0	
$C_{ACED} = \{A,C,E,D\}$	-			0

En base a estos cálculos, la matriz de distancias obtenida sería la siguiente:

	$C_B = \{B\}$	$C_{ACED} = \{A,C,E,D\}$
$C_B = \{B\}$	0	0.64
$C_{ACED} = \{A,C,E,D\}$	0.64	0

Por tanto en el **Nivel 3** tendremos dos clústeres:

$$C_B = \{B\}$$

$$C_{ACED} = \{A, C, E, D\}$$

Como no todas las instancias forman parte del mismo clúster se repite el proceso.

En la **cuarta iteración** observamos que ya sólo quedan dos clústeres: C_B = {B} y C_{ACED} = {A, C, E, D}. Por lo tanto unimos estos dos clústeres en el clúster C_{ACEDB} = {A, C, E, D, B}.

Entonces en el **Nivel 4** tendremos un único clúster:

$$C_{ACEDB} = \{A, C, E, D, B\}$$

Finalizamos el algoritmo porque todas las instancias forman parte del mismo clúster.

Por último representamos la estructura jerárquica de clústeres obtenida en el siguiente dendograma: