Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

Actividades resueltas

Algoritmo de *clustering* jerárquico

Descripción de la actividad

Una universidad online tiene la siguiente base de datos con cinco registros correspondientes a las notas de los alumnos en dos asignaturas de un máster.

	Asignatura 1	Asignatura 2
A1	1.5	2.0
A2	3.0	4.0
A3	5.5	7.0
A4	3.5	5.0
A5	4.5	5.0

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

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

El primer paso es calcular la matriz de distancias entre las diferentes instancias. Para ello se utiliza la distancia Euclídea y se calculan las distancias entre cada par de elementos (instancias).

Nota: La matriz es simétrica y en su diagonal tendrá como valores ceros, por lo que sólo en necesario calcular una parte de la matriz.

	A1 (1.5, 2.0)	A2 (3.0, 4.0)	A3 (5.5, 7.0)	A4 (3.5, 5.0)	A5 (4.5, 5.0)
A1 (1.5, 2.0)	0	$\sqrt{1.5^2 + 2^2} = \sqrt{6.25} = 2.50$	$\sqrt{4^2 + 5^2} = \sqrt{41} = 6.40$	$\sqrt{2^2 + 3^2} = \sqrt{13} = 3.61$	$\sqrt{3^2 + 3^2} = \sqrt{18} = 4.24$
A2 (3.0, 4.0)	-	0	$\sqrt{2.5^2 + 3^2} = \sqrt{15.25} = 3.91$	$\sqrt{0.5^2 + 1^2} = \sqrt{1.25} = 1.12$	$\sqrt{1.5^2 + 1^2} = \sqrt{3.25} = 1.80$
A3 (5.5, 7.0)	-	-	0	$\sqrt{2^2 + 2^2} = \sqrt{8} = 2.83$	$\sqrt{1^2 + 2^2} = \sqrt{5} = 2.24$
A4 (3.5, 5.0)	-	-	-	0	$\sqrt{1^2 + 0^2} = 1.00$
A5 (4.5, 5.0)	-	-	-	-	0

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

	A1	A2	A3	A4	A5
A1	0	2.50	6.40	3.61	4.24
A2	2.50	0	3.91	1.12	1.80
A3	6.40	3.91	0	2.83	2.24
A4	3.61	1.12	2.83	0	1.00
A5	4.24	1.80	2.24	1.00	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:

$$C1 = \{A1\}$$
$$C2 = \{A2\}$$
$$C3 = \{A3\}$$
$$C4 = \{A4\}$$
$$C5 = \{A5\}$$

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 A4 y A5. Estas dos instancias se encuentran a distancia 1.

Por lo tanto se unen los clústeres C4 = {A4} y C5 = {A5} en el nuevo clúster C4_5 = {A4, A5}.

Finalmente se recalcula la matriz de distancias. Para ello se eliminan las filas y columnas correspondientes a C4 y C5, se añade una nueva fila y una nueva columna para C4_5 y se recalculan las distancias mediante la medida de enlace sencillo. Como se está utilizando enlace sencillo la distancia entre el clúster C4_5 y el clúster C_i será el mínimo de la distancia entre el clúster C_i y el clúster C4 y la distancia entre el clúster C_i y el clúster C5.

	C1 ={A1}	C2 ={A2}	C3 ={A3}	C4 ={A4}	C5 ={A5}	C4_5 ={A4, A5}
C1 ={A1}	0	2.50	6.40	3.61	4.24	min{3.61, 4.24} =3.61
C2 ={A2}	2.50	0	3.91	1.12	1.80	min{1.12, 1.80} =1.12
C3 ={A3}	6.40	3.91	0	2.83	2.24	min{2.83, 2.24} =2.24
C4 ={A4}	3.61	1.12	2.83	0	1.00	
C5 ={A5}	4.24	1.80	2.24	1.00	0	
C4_5 ={A4, A5}	-	-	-			0

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

	C1 ={A1}	C2 ={A2}	C3 ={A3}	C4_5 ={A4, A5}
C1 ={A1}	0	2.50	6.40	3.61
C2 ={A2}	2.50	0	3.91	1.12
C3 ={A3}	6.40	3.91	0	2.24
C4_5 ={A4, A5}	3.61	1.12	2.24	0

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

C1 = {A1}

C2 = {A2}

C3 = {A3}

C4_5 = {A4, A5}

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 C2 y el clúster C4_5 que se encuentran a distancia 1.12.

Por lo tanto se unen los clústeres C2 = {A2} y C4_5 = {A4, A5} en el nuevo clúster C4_5_2 = {A4, A5, A2}.

Finalmente se recalcula la matriz de distancias. Para ello se eliminan las filas y columnas correspondientes a C2 y C4_5, se añade una nueva fila y una nueva columna para C4_5_2 y se recalculan las distancias mediante la medida de enlace sencillo.

	C1 ={A1}	C2 ={A2}	C3 ={A3}	C4_5 ={A4, A5}	C4_5_2 ={A4, A5, A2}
C1 ={A1}	0	2.50	6.40	3.61	min{2.50, 3.61} =2.50
C2 ={A2}	2.50	0	3.91	1.12	
C3 ={A3}	6.40	3.91	0	2.24	min{3.91, 2.24} =2.24
C4_5 ={A4, A5}	3.61	1.12	2.24	0	
C4_5_2 ={A4, A5, A2}	-		-		0

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

	C1 ={A1}	C3 ={A3}	C4_5_2 ={A4, A5, A2}
C1 ={A1}	0	6.40	2.50
C3 ={A3}	6.40	0	2.24
C4_5_2 ={A4, A5, A2}	2.50	2.24	0

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

$$C1 = \{A1\}$$

$$C3 = \{A3\}$$

$$C4\_5\_2 = \{A4, A5, A2\}$$

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 C3 y el clúster C4_5_2 que se encuentran a distancia 2.24.

Por lo tanto se unen los clústeres C3 = {A3} y C4_5_2 = {A4, A5, A2} en el nuevo clúster C4_5_2_3 = {A4, A5, A2, A3}.

Finalmente se recalcula la matriz de distancias. Para ello se eliminan las filas y columnas correspondientes a C3 y C4_5_2, se añade una nueva fila y una nueva columna para C4_5_2_3 y se recalculan las distancias mediante la medida de enlace sencillo.

	C1 ={A1}	C3 ={A3}	C4_5_2 ={A4, A5, A2}	C4_5_2_3 ={A4, A5, A2, A3}
C1 ={A1}	0	6.40	2.50	min{6.40, 2.50} =2.50
C3 ={A3}	6.40	0	2.24	
C4_5_2 ={A4, A5, A2}	2.50	2.24	0	
C4_5_2_3 ={A4, A5, A2, A3}	-			0

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

	C1 ={A1}	C4_5_2_3 ={A4, A5, A2, A3}
C1 ={A1}	0	2.50
C4_5_2_3 ={A4, A5, A2, A3}	2.50	0

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

$$C1 = \{A1\}$$

$$C4\_5\_2\_3 = \{A4, A5, A2, A3\}$$

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

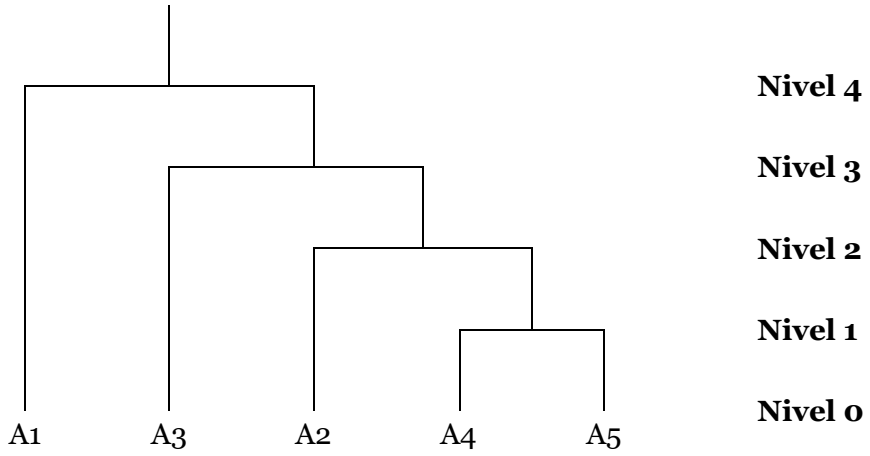
En la **cuarta iteración** observamos que ya sólo quedan dos clústeres: C1 = {A1} y C4_5_2_3 = {A4, A5, A2, A3}. Por lo tanto unimos estos dos clústeres en el clúster C4_5_2_3_1 = {A4, A5, A2, A3, A1}.

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

$$C4\_5\_2\_3\_1 = \{A4, A5, A2, A3, A1\}$$

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

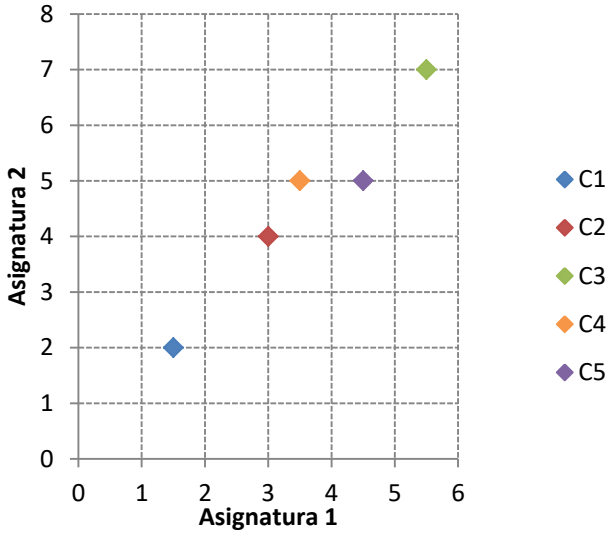
Por último representamos la estructura jerárquica de clústeres obtenida en el siguiente dendograma:



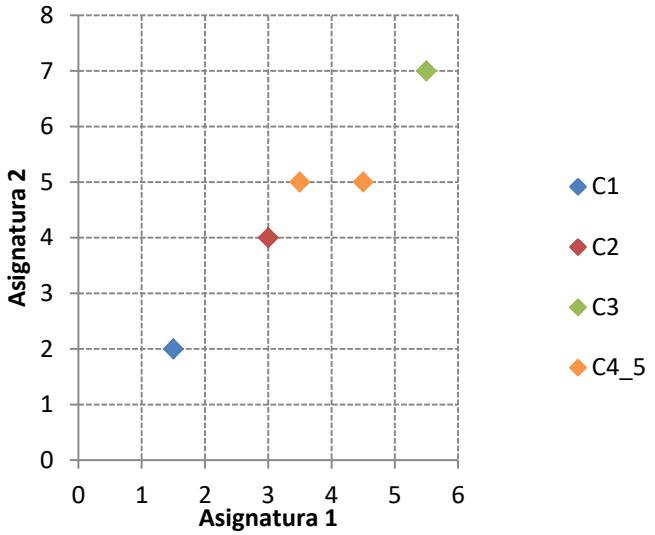
Anexo: Representación gráfica

Aunque no se pedía en el ejercicio, en los siguientes gráficos se representan los clústeres para cada nivel.

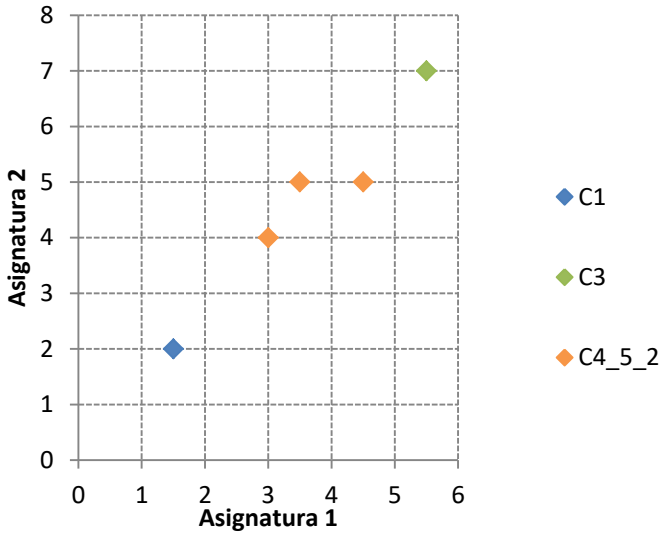
Nivel 0



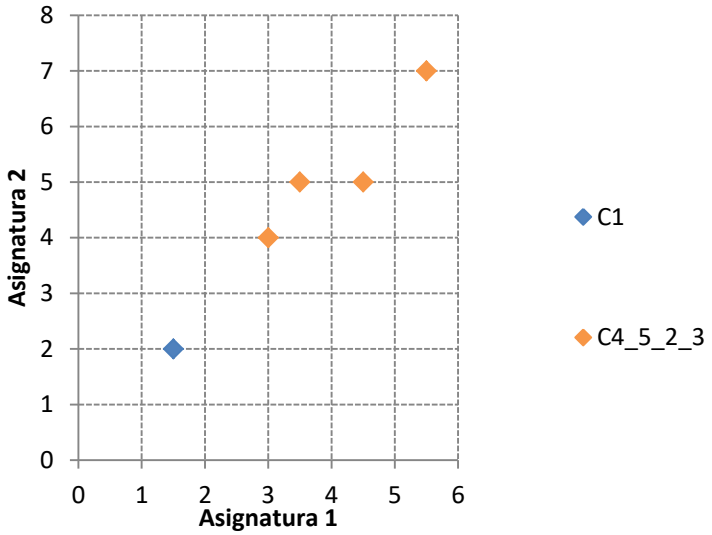
Nivel 1



Nivel 2



Nivel 3



Nivel 4