

Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

## Actividades resueltas

### Weka: Cobweb – Clústering jerárquico

#### Descripción de la actividad

Se ha obtenido la siguiente salida al aplicar Weka a un conjunto de datos.

=== Clustering model (full training set) ===

Number of merges: 130

Number of splits: 98

Number of clusters: 8

```
node 0 [768]
| node 1 [543]
| | leaf 2 [96]
| node 1 [543]
| | node 3 [254]
| | | leaf 4 [148]
| | node 3 [254]
| | | leaf 5 [106]
| node 1 [543]
| | leaf 6 [193]
node 0 [768]
| leaf 7 [225]
```

Time taken to build model (full training data) : 0.45 seconds

=== Model and evaluation on training set ===

Clustered Instances

```
2  111 ( 14%)
4  150 ( 20%)
5  108 ( 14%)
6  171 ( 22%)
7  228 ( 30%)
```

Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

Contesta a las siguientes cuestiones:

- ¿Qué algoritmo se ha aplicado? ¿Qué tipo de agrupamiento proporciona este algoritmo? ¿A qué categoría de aprendizaje pertenece este algoritmo?
- ¿Cuántos clústeres se han generado? ¿Cuántas instancias de los datos de entrenamiento conforman cada uno de los clústeres? Y en el proceso de evaluación del modelo ¿cuántas instancias conforman cada uno de los clústeres?
- Si se visualizaran los clústeres mediante la opción ‘Visualize clusters assignments’ de Weka en una gráfica que presentara dos de los atributos de entrada ¿qué clústeres observarías? ¿por qué?

### Resolución de la actividad

- ¿Qué algoritmo se ha aplicado?  
Cobweb

¿Qué tipo de agrupamiento proporciona este algoritmo?  
Agrupamiento o clustering jerárquico

*Explicación de la respuesta:*

El modelo obtenido que parece un árbol con nodos y hojas nos permite deducir que se trata de un algoritmo de clustering jerárquico. Además la información sobre el número de uniones (number of merges) y el número de divisiones (number of splits) nos indica que en la generación del modelo se ha combinado el comportamiento aglomerativo y divisorio de los algoritmos de clustering jerárquico.

¿A qué categoría de aprendizaje pertenece este algoritmo?  
Aprendizaje no supervisado

Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

*Explicación de la respuesta:*

Todos algoritmos de clústering, también Cobweb, pertenecen a la categoría de aprendizaje no supervisado porque a partir de las instancias que conforman los datos de entrenamiento se pueden identificar agrupamientos con instancias similares entre sí, es decir instancias que tienen valores similares para sus atributos.

- ¿Cuántos clústeres se han generado?  
8 clústeres

*Explicación de la respuesta:*

Nos viene indicado en la descripción del modelo Number of clusters: 8. Hay que destacar que en los agrupamientos jerárquicos una instancia puede formar parte de varios clústeres. De hecho los clústeres forman una estructura jerárquica con diferentes niveles y se pueden representar en un dendograma. Sin embargo Weka representa la jerarquía de clústeres en forma de árbol donde los clústeres son nodos hoja (leaf) y nodos (node). El modelo obtenido es:

```

node 0 [768]
| node 1 [543]
| | leaf 2 [96]
| node 1 [543]
| | node 3 [254]
| | | leaf 4 [148]
| | node 3 [254]
| | | leaf 5 [106]
| node 1 [543]
| | leaf 6 [193]
node 0 [768]
| leaf 7 [225]
```

En esta jerarquía de clústeres el Clúster 2 (leaf 2), el Clúster 4 (leaf 4), el Clúster 5 (leaf 5), el Clúster 6 (leaf 6) y el Clúster 7 (leaf 7) son los clústeres del nivel 0 del dendograma. El Clúster 3 (node 3) estaría en el nivel 1 del dendograma y aglomera el Clúster 4 y el Clúster 5. El Clúster 1 (node 1) estaría en el nivel 2 del dendograma y aglomera el Clúster 2, el Clúster 3 y el Clúster 6. Por último el Clúster 0 (node 0) estaría en el nivel 3 del dendograma y aglomera el Clúster 1 y el Clúster 7.

Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

¿Cuántas instancias de los datos de entrenamiento conforman cada uno de los clústeres?

Clúster 0: 768 instancias

Clúster 1: 543 instancias

Clúster 2: 96 instancias

Clúster 3: 254 instancias

Clúster 4: 148 instancias

Clúster 5: 106 instancias

Clúster 6: 193 instancias

Clúster 7: 225 instancias

*Explicación de la respuesta:*

Para proporcionar el número de instancias de los datos de entrenamientos que conforman cada uno de los clústeres al finalizar el proceso de aprendizaje de los clústeres se deben considerar los valores que aparecen entre corchetes en la información sobre el modelo generado y que indican el número de instancias de los datos de entrenamiento que forman el clúster correspondiente.

Concretamente:

```

node 0 [768]
| node 1 [543]
| | leaf 2 [96]
| | node 1 [543]
| | node 3 [254]
| | | leaf 4 [148]
| | | node 3 [254]
| | | leaf 5 [106]
| | node 1 [543]
| | leaf 6 [193]
node 0 [768]
| leaf 7 [225]
```

Y en el proceso de evaluación del modelo ¿cuántas instancias conforman cada uno de los clústeres?

Clúster 0: 768 instancias

Clúster 1: 540 instancias

Clúster 2: 111 instancias

Clúster 3: 258 instancias

Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

Clúster 4: 150 instancias

Clúster 5: 108 instancias

Clúster 6: 171 instancias

Clúster 7: 228 instancias

*Explicación de la respuesta:*

Considerando el número de instancias del conjunto de datos de prueba que conforman cada uno de los clústeres al evaluar el modelo, Weka presenta solamente el número de instancias que caen en los clústeres del nivel o del dendograma (nodos hoja del árbol):

Clustered Instances

2 111 ( 14%)

4 150 ( 20%)

5 108 ( 14%)

6 171 ( 22%)

7 228 ( 30%)

Entonces para los clústeres de los niveles superiores del dendograma tenemos que calcular el número de instancias que los conforman sumando el número de instancias de los clústeres de los niveles inferiores.

Clúster 3:

150 instancias del Clúster 4 + 108 instancias del Clúster 5 = 258 instancias

Clúster 1:

111 instancias del Clúster 2 + 258 instancias del Clúster 3 + 171 instancias del Clúster 6 = 540 instancias

Clúster 0:

540 instancias del Clúster 1 + 228 instancias del Clúster 7 = 768 instancias

- Si se visualizaran los clústeres mediante la opción 'Visualize clusters assignments' de Weka en una gráfica que presentara dos de los atributos de entrada ¿qué clústeres observarías? ¿por qué?

El Clúster 2, el Clúster 4, el Clúster 5, el Clúster 6 y el Clúster 7 porque son los que pertenecen al nivel o del dendograma (nodos hoja en la representación que hace Weka). El resto de clústeres, Clúster 0, Clúster 1 y Clúster 3, y que pertenecen a los niveles superiores del dendograma, contienen las mismas instancias que conforman los clústeres del nivel o y no se representan en las

Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

gráficas que proporciona Weka. Esto se debe a que una instancia que pertenece al Clúster 2 también pertenece al Clúster 1 y al Clúster 0 y para facilitar su representación en la gráfica se pinta solamente del color del Clúster 2 (clúster de nivel 0 en el dendograma o nodo hoja en Weka).