Clustering en Weka

Laura Rodríguez Navas
rodrigueznavas@posgrado.uimp.es

July 10, 2020

En esta práctica se realiza un estudio acerca de la base de datos Iris. Esta BD se distribuye junto a la herramienta Weka.

1. Ejecuta el algoritmo SimpleKMeans usando la herramienta Weka con las distancias Euclídea y Manhattan.

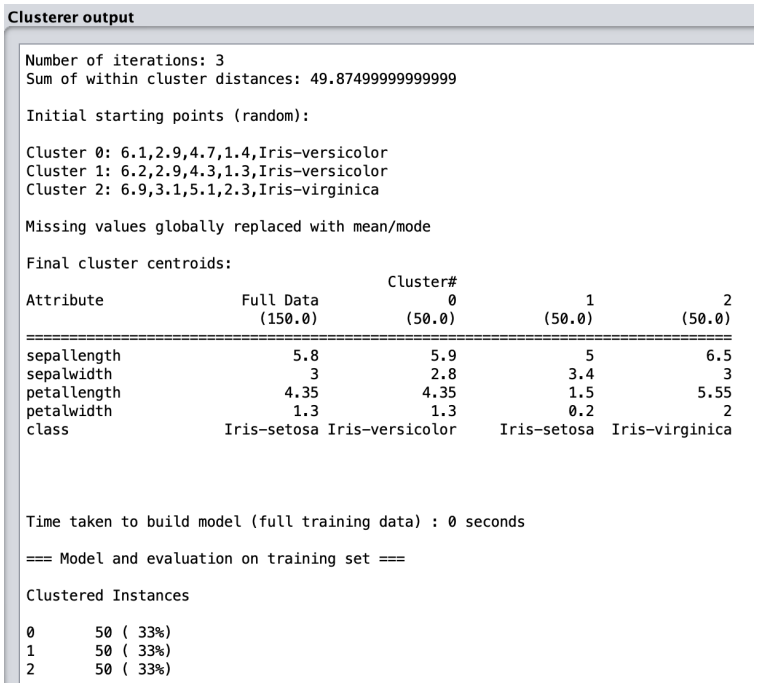
   La BD está formada por 4 variables descriptivas y una variable clase. No se aplica preprocesamiento de datos ya que las variables descriptivas son numéricas y no existen valores perdidos en la BD. Además, como la variable de clase puede tomar tres valores, elegimos que el valor de k sea igual a 3.

Clusterer output

```
Number of iterations: 3
Within cluster sum of squared errors: 7.817456892309574

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4,Iris-versicolor
Cluster 1: 6.2,2.9,4.3,1.3,Iris-versicolor
Cluster 2: 6.9,3.1,5.1,2.3,Iris-virginica

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute               Full Data             0              1              2
                          (150.0)          (50.0)         (50.0)         (50.0)
================================================================================
sepallength                5.8433           5.936          5.006          6.588
sepalwidth                  3.054            2.77          3.418          2.974
petallength                3.7587            4.26          1.464          5.552
petalwidth                 1.1987           1.326          0.244          2.026
class                 Iris-setosa Iris-versicolor     Iris-setosa  Iris-virginica



Time taken to build model (full training data) : 0 seconds

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

Clustered Instances

0       50 ( 33%)
1       50 ( 33%)
2       50 ( 33%)
```

Figura 1: KMeans con distancia Euclídea.

   (a) ¿Cuántas instancias contiene cada grupo?
   En la ejecución del algoritmo KMeans con distancia Euclídea (ver Figura 1) se han formado tres grupos: 0, 1 y 2. Los 3 grupos contienen 50 instancias cada uno. En la ejecución del algoritmo KMeans con distancia Manhattan (ver Figura 2) también se han formado los grupos 0, 1 y 2, con el mismo número de instancias cada uno.



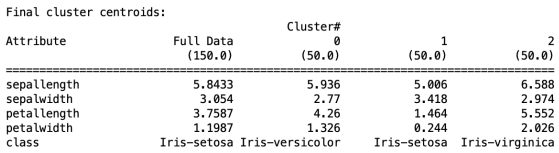
Clusterer output

```
Number of iterations: 3
Sum of within cluster distances: 49.87499999999999

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4,Iris-versicolor
Cluster 1: 6.2,2.9,4.3,1.3,Iris-versicolor
Cluster 2: 6.9,3.1,5.1,2.3,Iris-virginica

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute               Full Data             0             1             2
                          (150.0)       (50.0)        (50.0)        (50.0)
================================================================================
sepallength                   5.8           5.9             5           6.5
sepalwidth                      3           2.8           3.4             3
petallength                  4.35          4.35           1.5          5.55
petalwidth                    1.3           1.3           0.2             2
class                 Iris-setosa Iris-versicolor     Iris-setosa  Iris-virginica



Time taken to build model (full training data) : 0 seconds

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

Clustered Instances

0       50 ( 33%)
1       50 ( 33%)
2       50 ( 33%)
```

Figura 2: KMeans con distancia Manhattan.

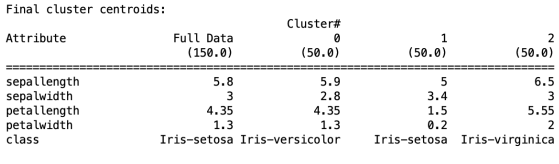
(b) ¿Cuáles son los centroides?

Si nos volvemos a fijar en la figura 1, podemos observar los centroides de la ejecución de Kmeans con distancia Euclídea, y en la figura 2, podemos observar los centroides de la ejecución de Kmeans con distancia Manhattan. Se muestran más detalladamente en las siguientes figuras:



```
Final cluster centroids:
                                        Cluster#
Attribute               Full Data             0             1             2
                          (150.0)       (50.0)        (50.0)        (50.0)
================================================================================
sepallength               5.8433         5.936         5.006         6.588
sepalwidth                 3.054          2.77         3.418         2.974
petallength               3.7587          4.26         1.464         5.552
petalwidth                1.1987         1.326         0.244         2.026
class                 Iris-setosa Iris-versicolor     Iris-setosa  Iris-virginica
```

Figura 3: KMeans centroides con distancia Euclídea.



```
Final cluster centroids:
                                        Cluster#
Attribute               Full Data             0             1             2
                          (150.0)       (50.0)        (50.0)        (50.0)
================================================================================
sepallength                   5.8           5.9             5           6.5
sepalwidth                      3           2.8           3.4             3
petallength                  4.35          4.35           1.5          5.55
petalwidth                    1.3           1.3           0.2             2
class                 Iris-setosa Iris-versicolor     Iris-setosa  Iris-virginica
```

Figura 4: KMeans centroides con distancia Manhattan.

(c) Analiza los centroides. ¿Hay algo destacable en esos centroides? ¿Están los centroides separados en el espacio? ¿Tienen componentes similares?

Los centroides resultantes de la ejecución del algoritmo KMeans con distancia Euclídea y los centroides resultantes de la ejecución del algoritmo KMeans con distancia Manhattan son muy parecidos. Tienen componentes muy similares. Este comportamiento nos podría indicar que en la BD no existen *outliers*, la distancia Manhattan se ve menos afectada por ellos (es más robusta), y al no presentar diferencias significantes con la distancia Euclídea, podría ser el motivo de tanta similitud.

En la figura 5 se puede observar que los centroides están separados en el espacio. Aunque los centroides de los grupos 0 y 1 están más cercanos entre ellos. El centroide del grupo 2 es el que está más alejado.

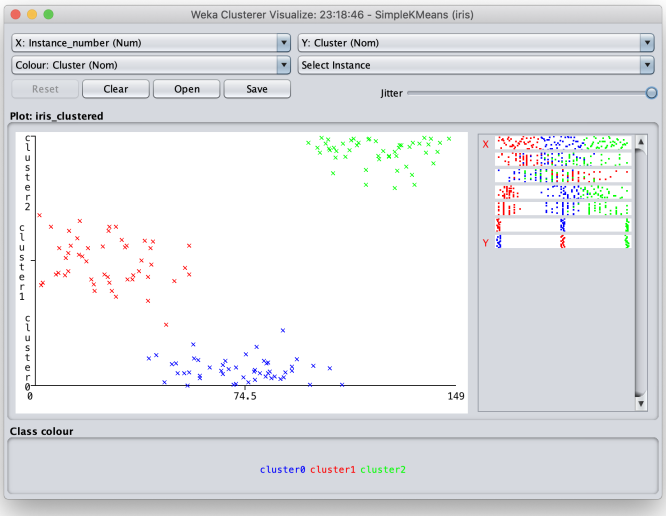


Figura 5: Representación de los centroides de KMeans con distancia Euclídea.

Nota: Solo se muestra la gráfica de los centroides resultantes de la ejecución del algoritmo KMeans con distancia Euclídea, porqué casi es igual a la gráfica de los centroides resultantes de la ejecución del algoritmo KMeans con distancia Manhattan.

2. Ejecuta el algoritmo HierarchicalClusterer con tipo de enlace completo y métrica de distancia euclídea, y visualice las gráficas de los puntos agrupados. ¿Alguno de ellas produce grupos bien diferenciados y con fronteras claras?

   Nota: Compara que el eje X instance_number y el eje Y vaya variando y muestra cada una de las variables (debes adjuntar las imágenes).

   Como se ha comentado anteriormente, la variable de clase puede tomar tres valores, así que volvemos a elegir k igual a 3 para la ejecución del algoritmo (ver Figura 6).

   Las gráficas donde el eje Y es igual a las variables *petallenght* i *petalwidth* muestran grupos bien diferenciados y con las fronteras claras.

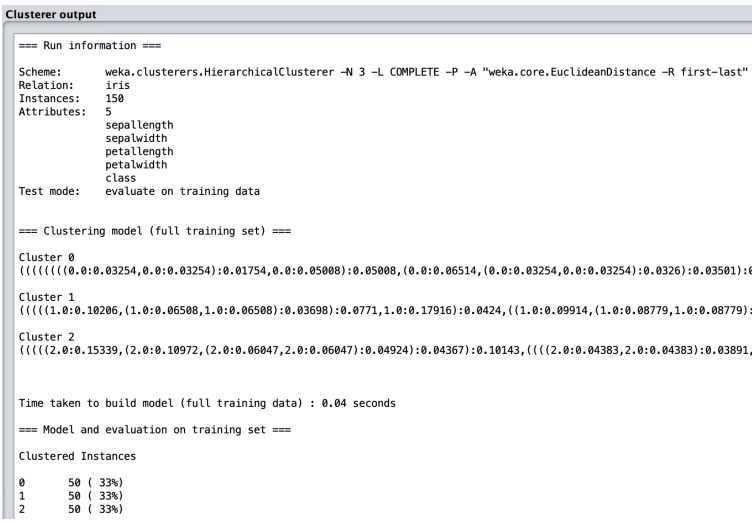


Figura 6: Hierarchical Clustering con distancia Euclídea.

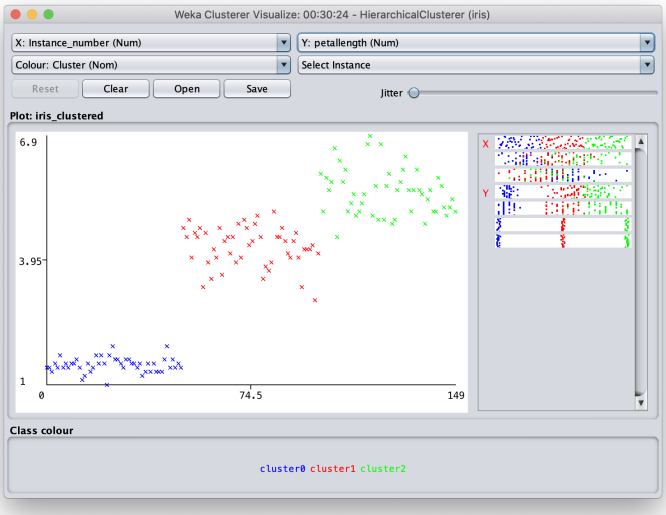


Figura 7: X instance_number con Y *petallenght*.

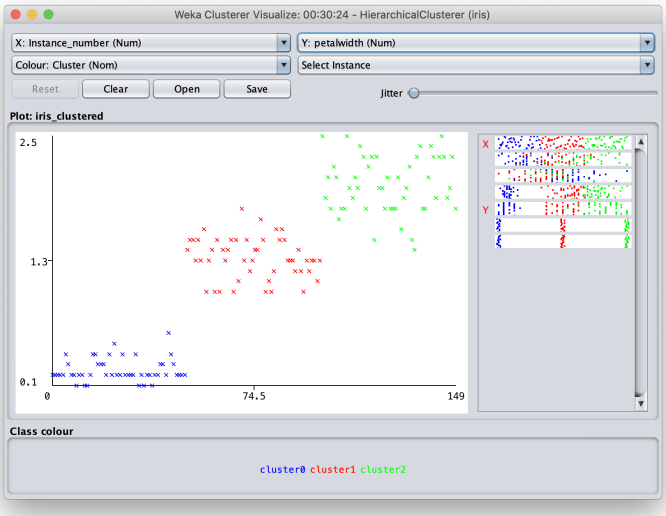


Figura 8: X instance_number con Y *petalwidth*.

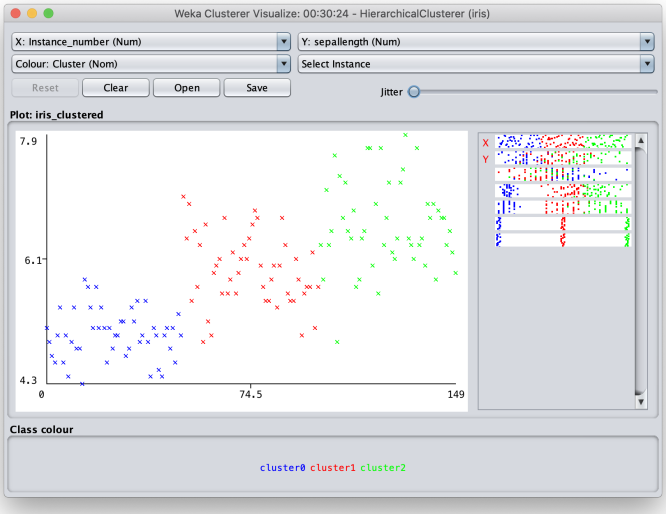


Figura 9: X instance_number con Y *sepalleght*.

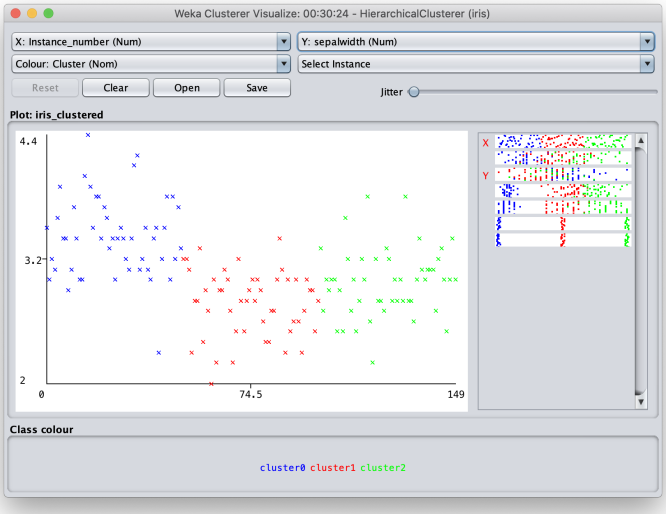


Figura 10: X instance_number con Y *sepalwidth*.