Clustering en Weka

Laura Rodríguez Navas
rodrigueznavas@posgrado.uimp.es

July 11, 2020

En esta práctica vamos a realizar un estudio acerca de los datos de Iris. Esta BD se distribuye junto a la herramienta Weka. Toda la información sobre esta BD se puede encontrar en el siguiente enlace: `https://archive.ics.uci.edu/ml/datasets/Iris`. La BD contiene 4 variables descriptivas y una variable clase. Dado que no contiene valores perdidos, no deberemos de eliminarlos antes de continuar el análisis. Además, como las variables descriptivas son numéricas no deberemos de transformarlas.

Debido a que los algoritmos de clustering se basan en el cómputo de valores de distancias entre los grupos, es ventajoso tener todas las variables en la misma escala para el cálculo de las distancias entre grupos. Así que, vamos a convertir todas las variables descriptivas a una misma escala. Para ello, vamos a utilizar un filtro de atributo de tipo no supervisado llamado *Standardize* (ver Figura 1), y vamos a ignorar la variable clase durante el proceso.

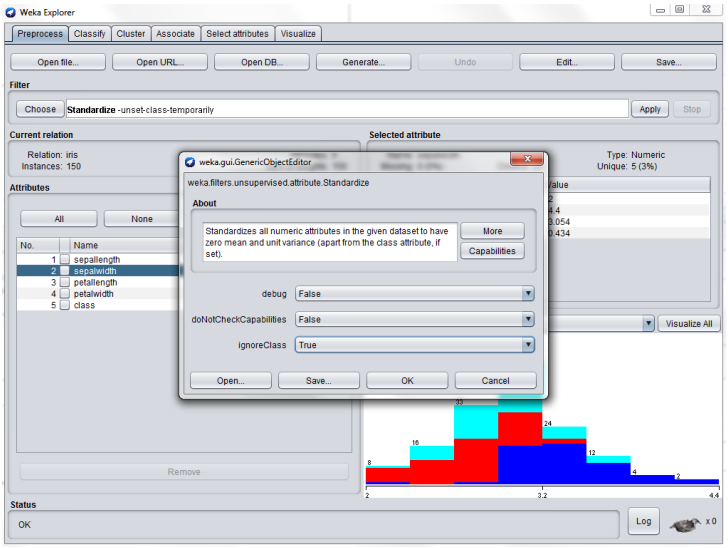


Figura 1: Filtro *Standardize*.

La variable clase no se debe tener en cuenta a la hora de realizar el clustering, ya que estamos haciendo un aprendizaje no supervisado, y por tanto la vamos a eliminar. Para ello, vamos a utilizar un filtro de atributo no supervisado llamado *Remove* (ver Figura 2).

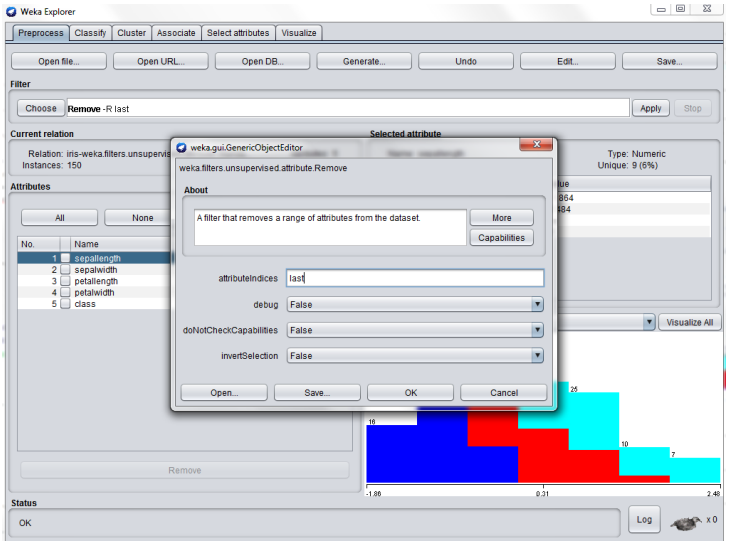


Figura 2: Filtro *Remove*.

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

```
kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 6.998114004826762

Initial starting points (random):

Cluster 0: 0.309959,-0.355171,0.533509,0.263815
Cluster 1: 0.430722,-0.355171,0.306805,0.132781
Cluster 2: 1.276066,0.10609,0.760212,1.443121

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute     Full Data         0          1          2
                (150.0)     (61.0)     (50.0)     (39.0)
============================================================
sepallength          -0     0.0546    -1.0112     1.211
sepalwidth           -0    -0.7295     0.8395     0.0647
petallength          -0     0.3616    -1.3005     1.1017
petalwidth           -0     0.2874    -1.2509     1.1542



Time taken to build model (full training data) : 0.01 seconds

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

Clustered Instances

0       61 ( 41%)
1       50 ( 33%)
2       39 ( 26%)
```

(a) con distancia Euclídea.

```
kMeans
======

Number of iterations: 5
Sum of within cluster distances: 47.77942561205271

Initial starting points (random):

Cluster 0: 0.309959,-0.355171,0.533509,0.263815
Cluster 1: 0.430722,-0.355171,0.306805,0.132781
Cluster 2: 1.276066,0.10609,0.760212,1.443121

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute     Full Data         0          1          2
                (150.0)     (62.0)     (50.0)     (38.0)
============================================================
sepallength     -0.0523     0.0684    -1.0184     1.0345
sepalwidth      -0.1245    -0.5858     0.798     -0.1245
petallength      0.3351     0.4202    -1.2801     1.0719
petalwidth       0.1328     0.2638    -1.3086     1.1811



Time taken to build model (full training data) : 0.01 seconds

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

Clustered Instances

0       62 ( 41%)
1       50 ( 33%)
2       38 ( 25%)
```

(b) con distancia Manhattan.

Figura 3: SimpleKMeans.

Nota: Como la variable de clase puede tomar 3 valores diferentes, hemos elegido que el valor de k sea igual a 3.

Si nos fijamos en el valor de SSE, vemos que el primer clustering con distancia Euclídea ha funcionado mejor, ya que el valor de SSE es pequeño (6.998). En cambio, el clustering con distancia Manhattan no ha funcionado muy bien, ya que el valor de SSE es muy alto (47.779). Para mejorar el segundo clustering, tendríamos que augmentar el valor de k.

(a) ¿Cuántas instancias contiene cada grupo?

En la Figura 3 podemos ver las instancias que contienen cada grupo (0, 1 y 2), que se resumen en la siguiente Tabla 1.

Distancia	Grupo 0	Grupo 1	Grupo 2
Euclídea	61	50	39
Manhattan	62	50	38

Tabla 1: Instancias de cada grupo en SimpleKMeans.

(b) ¿Cuáles son los centroides?

Si nos volvemos a fijar en la figura 3, podemos observar los centroides de cada ejecución de SimpleKMeans, que se muestran más detalladamente en la siguiente figura.

```
Final cluster centroids:
                        Cluster#
Attribute     Full Data         0         1         2
                (150.0)    (61.0)    (50.0)    (39.0)
============================================================
sepallength          -0    0.0546   -1.0112     1.211
sepalwidth           -0   -0.7295    0.8395    0.0647
petallength          -0    0.3616   -1.3005    1.1017
petalwidth           -0    0.2874   -1.2509    1.1542
```

(a) con distancia Euclídea.

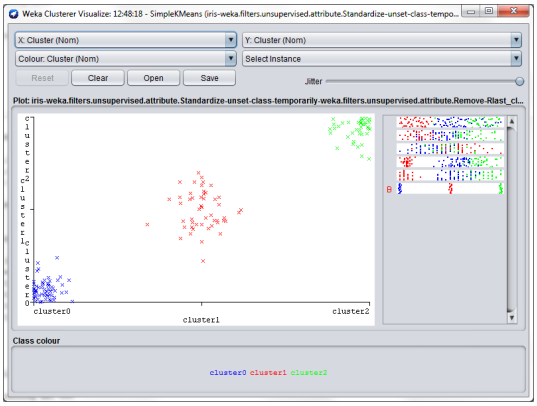
```
Final cluster centroids:
                        Cluster#
Attribute     Full Data         0         1         2
                (150.0)    (62.0)    (50.0)    (38.0)
============================================================
sepallength    -0.0523    0.0684   -1.0184    1.0345
sepalwidth     -0.1245   -0.5858     0.798   -0.1245
petallength     0.3351    0.4202   -1.2801    1.0719
petalwidth      0.1328    0.2638   -1.3086    1.1811
```

(b) con distancia Manhattan.

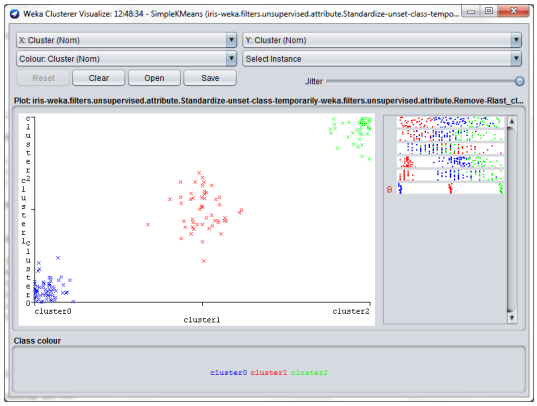
Figura 4: Centroides en SimpleKMeans.

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

En general, vemos que hay bastante cohesión entre los centroides, eso quiere decir que tienen componentes muy similares. Por ejemplo, el grupo 0 presenta mucha cohesión. También vemos que existe mucha separación entre los centroides. A continuación, podemos observar visualmente la separación entre los centroides (ver Figura 5).



(a) con distancia Euclídea.



(b) con distancia Manhattan.

Figura 5: Centroides en SimpleKMeans.

2. Ejecuta el algoritmo HierarchicalClusterer con tipo de enlace completo y métrica de distancia Euclídea, y visualiza las gráficas de los puntos agrupados. ¿Alguna 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).

Nota: Volvemos a elegir para la ejecución del algoritmo que el valor de k sea igual a 3.

El resultado de la ejecución del algoritmo HierarchicalClusterer se puede observar en la siguiente figura:

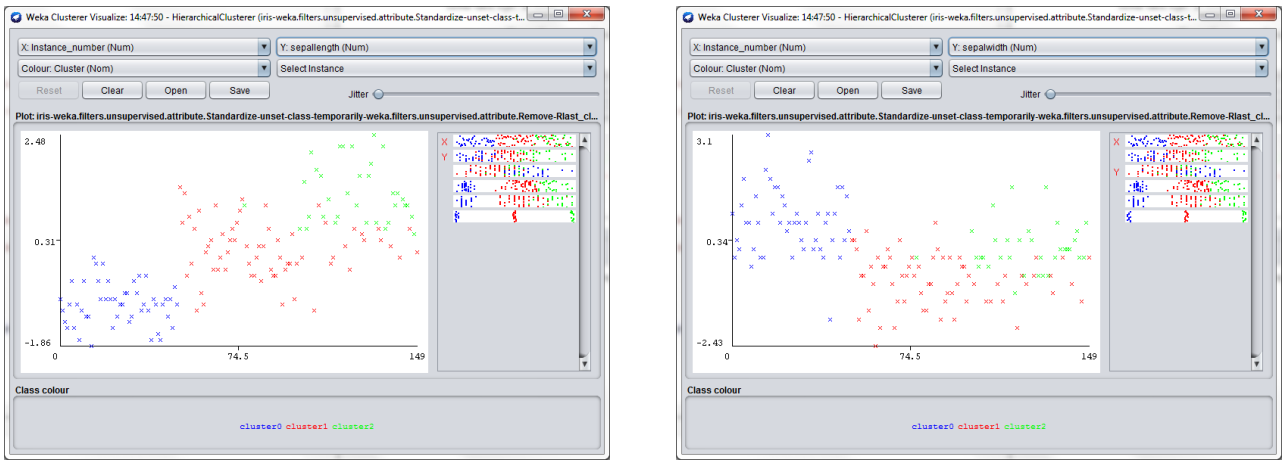
```
=== Run information ===

Scheme:       weka.clusterers.HierarchicalClusterer -N 3 -L COMPLETE -P -A "weka.core.EuclideanDistance -R first-last"
Relation:     iris-weka.filters.unsupervised.attribute.Standardize-unset-class-temporarily-weka.filters.unsupervised.attribute.Remove-Rlast
Instances:    150
Attributes:   4
              sepallength
              sepalwidth
              petallength
              petalwidth
Test mode:    evaluate on training data


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

Cluster 0
(((((((-1.3085928194379581:0.03254,-1.3085928194379581:0.03254):0.01754,-1.3085928194379581:0.05008):0.05008,(-1.3085928194379581:0.06514,

Cluster 1
(((((0.263815107790562:0.10206,(0.39484910172627213:0.06508,0.39484910172627213:0.06508):0.03698):0.1195,((0.39484910172627213:0.09914,(0.1

Cluster 2
(((((0.6569170895976921:0.16107,(1.050019071404822:0.08504,1.050019071404822:0.08504):0.07603):0.02014,(0.7879510835334022:0.08946,(0.787951



Time taken to build model (full training data) : 0.04 seconds

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

Clustered Instances

0       50 ( 33%)
1       66 ( 44%)
2       34 ( 23%)
```

Figura 6: HierarchicalClusterer con distancia Euclídea.

La ejecución del algoritmo HierarchicalClusterer ha formado el grupo 0 con 50 instancias, el grupo 1 con 66 instancias, y el grupo 2 con 34 instancias. Aunque el valor de k es el mismo que en la ejecución del algoritmo SimpleKMeans, no se obtiene el mismo resultado (ver Tabla 1).

A continuación se visualizan ls gráficas de los puntos agrupados para cada variable descriptiva del eje Y con el eje X instance_number.

- *sepallength* y *sepalwidth*

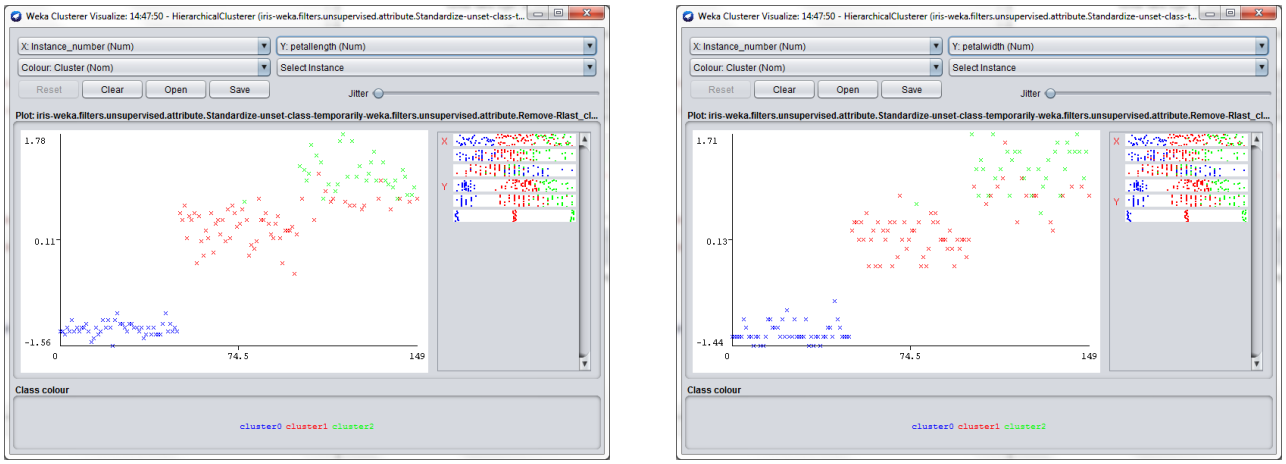


(a) con Y *sepallength*.

(b) con Y *sepalwidth*.

Figura 7: X instance_number con *sepallength* y *sepalwidth*.

- *petallength* y *petalwidth*



(a) con Y *petallength*.

(b) con Y *petalwidth*.

Figura 8: X instance_number con *petallength* y *petalwidth*.

En ninguna de ellas, todos los grupos están totalmente diferenciados y con fronteras claras.

En las gráficas de la Figura 7 los grupos no se han agrupado correctamente.

En las gráficas de la Figura 8 los grupos 1 y 2 no se han agrupado correctamente. Sí que se ha agrupado correctamente el grupo 0.