

Trabajo 9

kMeans

Facultad De Ingeniería, Universidad De Cuenca
TEXT MINING

Freddy L. Abad L.

freddy.abadl@ucuenca.edu.ec

En esta práctica se verá los elementos necesarios para generar un proceso de agrupamiento usando K-means. El objetivo es que este proceso puede ser aplicado a cualquier caso de prueba que requiera ejecutar este tipo de agrupamiento. En primera instancia se aplicará el proceso para un caso de uso, pero más adelante se ejecutará la misma tarea para otro escenario, indicando los pasos necesarios para reutilizar el proceso creado en una nueva base de datos. En general un proceso de agrupamiento basado en k-means requiere cuatro operadores.

Retrieve or Read CSV: Este operador permitirá leer datos desde una fuente. En este caso se usará la BD IRIS <https://archive.ics.uci.edu/ml/datasets/iris>. Esta fuente de información es quizás la base de datos más conocida que se encuentra en la literatura de reconocimiento de patrones.

Proceso de Lectura de Archivo Datos



Configuración de proceso Read Csv

Parameters

Read CSV

Import Configuration Wizard...

csv file: 9\datosPractica9.csv

column separators: ,

☐ trim lines

☒ use quotes

quotes character: "

escape character: \

☐ skip comments

starting row: 1

La base de datos contiene la siguiente información:

1. longitud del sépalos en cm
2. ancho del sépalos en cm

3. longitud del pétalo en cm
4. ancho del pétalo en cm
5. clase:
 - Iris Setosa
 - Iris Versicolor
 - Iris Virginica

La idea en esta práctica es tratar de buscar agrupamiento de las flores considerando las variables de la 1 a la 4, sin considerar por supuesto el atributo 5 que representan la clase de flor.

Descripción de Archivo Fuente

```
datosPractica9.csv x
1 | longsepalo, anchosepalo, longpetalo, anchopetalo, clase
2 | 5.1, 3.5, 1.4, 0.2, Iris-setosa
3 | 4.9, 3.0, 1.4, 0.2, Iris-setosa
4 | 4.7, 3.2, 1.3, 0.2, Iris-setosa
5 | 4.6, 3.1, 1.5, 0.2, Iris-setosa
```

Seleccionar atributos de archivos, se elimina la última columna por razones didácticas, y luego de comprobación



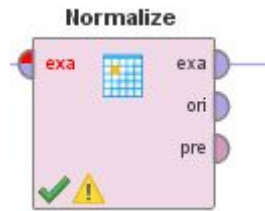
Salida de Proceso

| Row No. | longsepalo | anchosepalo | longpetalo | anchopetalo |
|---------|------------|-------------|------------|-------------|
| 7 | 4.600 | 3.400 | 1.400 | 0.300 |
| 8 | 5 | 3.400 | 1.500 | 0.200 |
| 9 | 4.400 | 2.900 | 1.400 | 0.200 |

Normalize: Este operador permite normalizar los datos. El objetivo de normalizar es poner a competir a todas las variables numéricas en los mismos términos. Dado que el agrupamiento en k-means se basa en distancias, por tanto lo ideal es que las variables sean aproximadamente las mismas. De modo que una variable que esta medida en una unidad muy grande no domine a otra que esta medida en una unidad más pequeña. El método usado para normalizar será Z-transformation el cual para cada una de las columnas le va a aplicar la media y lo va a dividir para la desviación estándar. Luego de aplicar este proceso

todas las medias serán 0 y las desviaciones estándar serán 1. Para observar estos valores use el botón Statistics una vez se ejecute el proceso. De esta manera podemos afirmar que todas las variables están en igualdad de condiciones

Normalización



Configuración de proceso Normalize (Z Transformation como método de Normalización)

Parameters

Normalize

☐ create view

attribute filter type

all

☐ invert selection

☐ include special attributes

method

Z-transformation

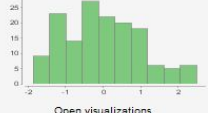
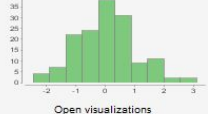
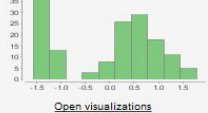
Salida Proceso

| ExampleSet (Normalize) | | | | |
|---|------------|-------------|------------|-------------|
| <div>Open in</div> <div> <div>Turbo Prep</div> <div>Auto Model</div> </div> | | | | |
| Row No. | longsepalo | anchosepalo | longpetalo | anchopetalo |
| 1 | -0.898 | 1.029 | -1.337 | -1.309 |
| 2 | -1.139 | -0.125 | -1.337 | -1.309 |
| 3 | -1.381 | 0.337 | -1.393 | -1.309 |
| 4 | -1.501 | 0.106 | -1.280 | -1.309 |
| 5 | -1.018 | 1.259 | -1.337 | -1.309 |
| 6 | -0.535 | 1.951 | -1.167 | -1.047 |
| 7 | -1.501 | 0.798 | -1.337 | -1.178 |
| 8 | -1.018 | 0.798 | -1.280 | -1.309 |

Estadísticas de la salida del proceso

| ExampleSet (Normalize) | | | | | | |
|------------------------|------|---------|------------|-------|----------------------------|--|
| Name | Type | Missing | Statistics | | Filter (4 / 4 attributes): | |
| longsepal | Real | 0 | Min | Max | Average | |
| | | | -1.864 | 2.484 | -0.000 | |
| anchosepal | Real | 0 | Min | Max | Average | |
| | | | -2.431 | 3.104 | -0.000 | |
| longpetal | Real | 0 | Min | Max | Average | |
| | | | -1.563 | 1.780 | -0.000 | |
| anchopetal | Real | 0 | Min | Max | Average | |
| | | | -1.440 | 1.705 | -0.000 | |

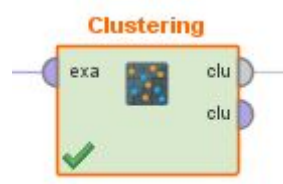
Estadística de Desviación Estándar y media de los datos

| ExampleSet (Normalize) | | | | | | |
|------------------------|------|---------|---|--|----------------------------|-----------|
| Name | Type | Missing | Statistics | | Filter (4 / 4 attributes): | |
| longsepal | Real | 0 |  | | Min | Max |
| | | | Open visualizations | | -1.864 | 2.484 |
| anchosepal | Real | 0 |  | | Min | Max |
| | | | Open visualizations | | -2.431 | 3.104 |
| longpetal | Real | 0 |  | | Min | Max |
| | | | Open visualizations | | -1.563 | 1.780 |
| anchopetal | Real | 0 |  | | Min | Max |
| | | | Open visualizations | | -1.440 | 1.705 |
| | | | | | Average | Deviation |
| | | | | | -0.000 | 1.000 |

K-means: Este operador permitirá ejecutar el proceso de agrupamiento. Es necesario indicar el valor de k, que indica el número de grupos que deseamos hallar. Es difícil conocer a priori cuántos grupos se pueden hallar en los datos, por tanto el objetivo es probar varias opciones hasta encontrar valores de similitud muy cercanos entre sí.

Luego de ejecutar este proceso se puede observar que al primer cluster pertenece 50 ítems y al segundo 100 ítems. En esta opción se puede observar además la tabla de centroides, lo cual indica las medias de las distintas variables una vez que las hemos separado en dos grupos. Al estar normalizadas las variables lo único que se puede indicar con estos valores es que por ejemplo la media del atributo sepal length es -1.011 y del segundo cluster es un valor positivo 0.506. No es posible interpretar directamente estos datos.

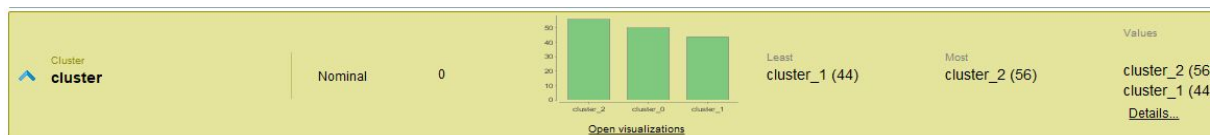
Proceso Clustering



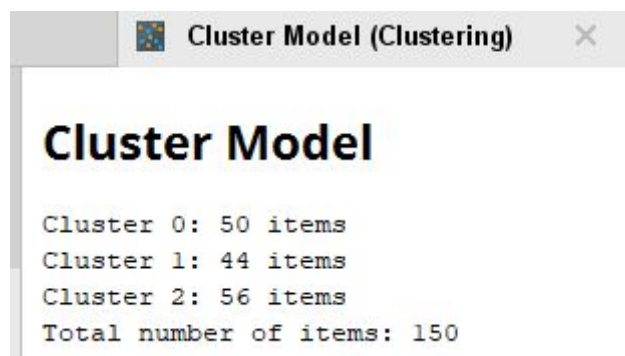
Salida proceso

| ExampleSet (Clustering) X | | | | | | |
|---------------------------------|----|-----------|------------|-------------|------------|-------------|
| Open in Turbo Prep Auto Model | | | | | | |
| Row No. | id | cluster | longsepalo | anchosepalo | longpetalo | anchopetalo |
| 1 | 1 | cluster_0 | -0.898 | 1.029 | -1.337 | -1.309 |
| 2 | 2 | cluster_0 | -1.139 | -0.125 | -1.337 | -1.309 |
| 3 | 3 | cluster_0 | -1.381 | 0.337 | -1.393 | -1.309 |
| 4 | 4 | cluster_0 | -1.501 | 0.106 | -1.280 | -1.309 |
| 5 | 5 | cluster_0 | -1.018 | 1.259 | -1.337 | -1.309 |
| 6 | 6 | cluster_0 | -0.535 | 1.951 | -1.167 | -1.047 |
| 7 | 7 | cluster_0 | -1.501 | 0.798 | -1.337 | -1.178 |
| 8 | 8 | cluster_0 | -1.018 | 0.798 | -1.337 | -1.309 |

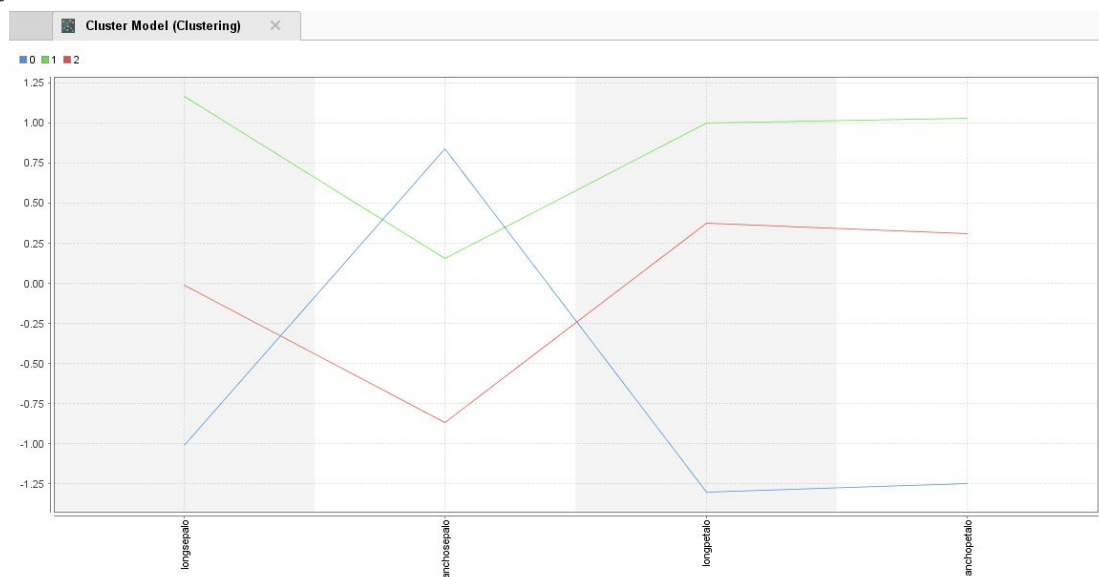
Estadística de la salida del proceso



Salida del proceso cluster



Gráfica del cluster de salida



Distancia entre centroides

Avg. within centroid distance

Avg. within centroid distance: 0.935

Distancia entre puntos del cluster 1

Avg. within centroid distance_cluster_0

Avg. within centroid distance_cluster_0: 0.963

Distancia entre puntos del cluster 2

Avg. within centroid distance_cluster_1

Avg. within centroid distance_cluster_1: 0.988

Distancia entre puntos del cluster 3

Avg. within centroid distance_cluster_2

Avg. within centroid distance_cluster_2: 0.867

Medida Davies Bouldin

Davies Bouldin

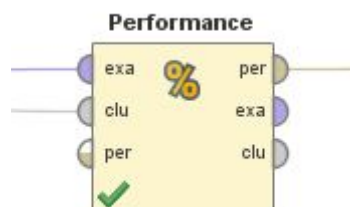
Davies Bouldin: 0.834

El gráfico usando la opción Plot permite diferenciar los valores entre los dos clúster. Por ejemplo aquí se puede visualizar que las flores rojas tienen los pétalos más anchos (petal width) y altos (petal height) que las flores azules.

Cluster Distance Performance: El cual va a permitir evaluar la calidad del agrupamiento, midiendo la distancia promedio de cada una de las flores en el grupo a su centroide de ese cluster. El centroide se convierte como en el representante de cada uno de los grupos.

Del operador de Clustering salen dos terminales:

El Cluster model debe conectarse al Cluster Model en el operador de Performance y el terminal clustered set debe conectarse al terminal example set.



Parameters

⚙️

Performance (Cluster Distance Performance)

main criterion

Avg. within centroid dista...

☐

main criterion only

☐

normalize

☒

maximize

Distancia entre centroides del cluster

Avg. within centroid distance

Avg. within centroid distance: 0.935

Distancia entre puntos del cluster 1

Avg. within centroid distance_cluster_0

Avg. within centroid distance_cluster_0: 0.963

Distancia entre puntos del cluster 2

Avg. within centroid distance_cluster_1

Avg. within centroid distance_cluster_1: 0.988

Distancia entre puntos del cluster 3

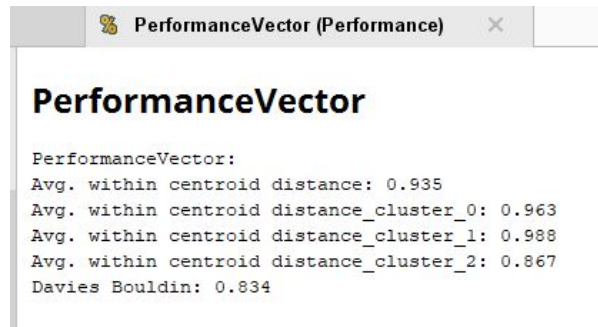
Avg. within centroid distance_cluster_2

Avg. within centroid distance_cluster_2: 0.867

Medida Davies Bouldin

Davies Bouldin

Davies Bouldin: 0.834

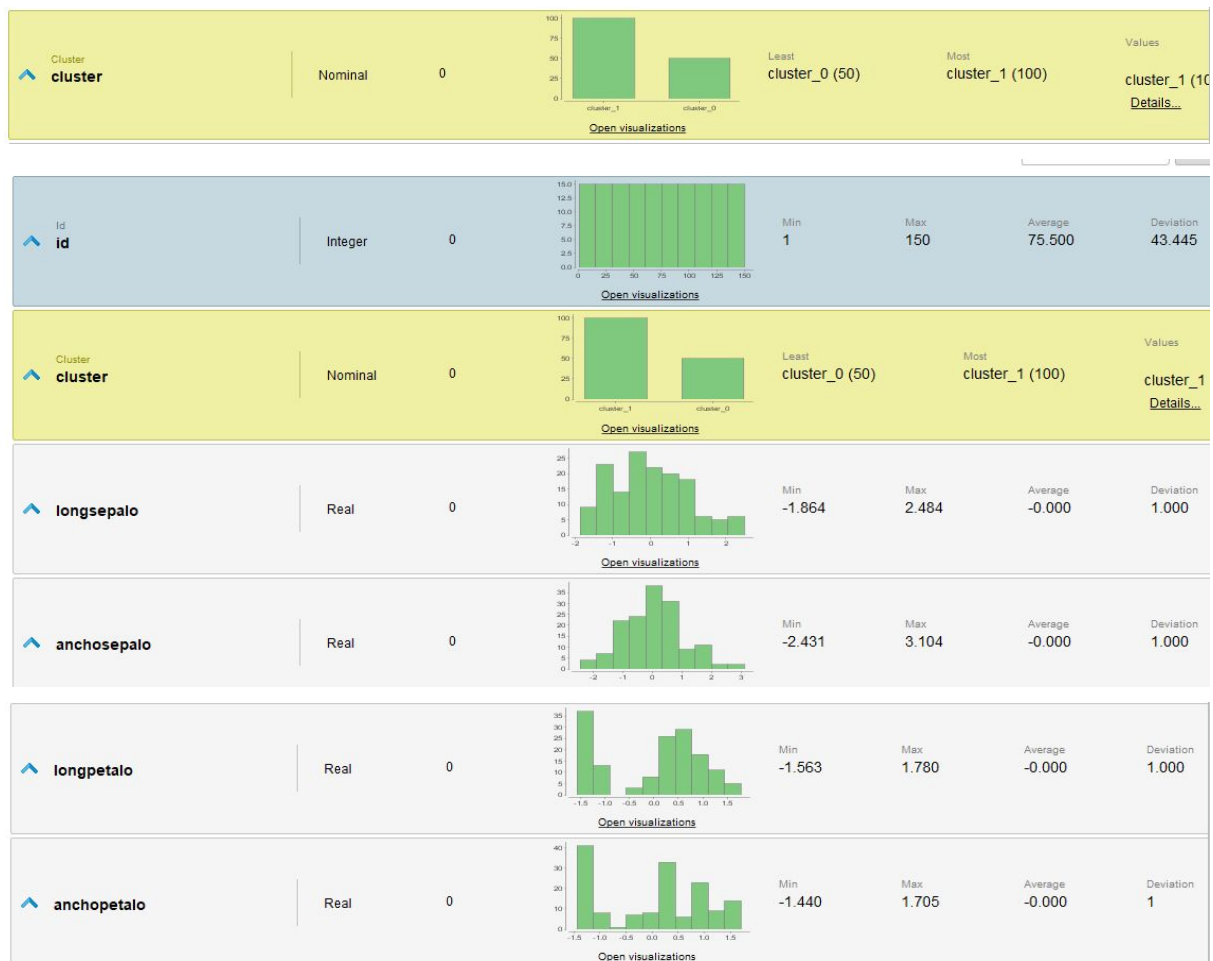


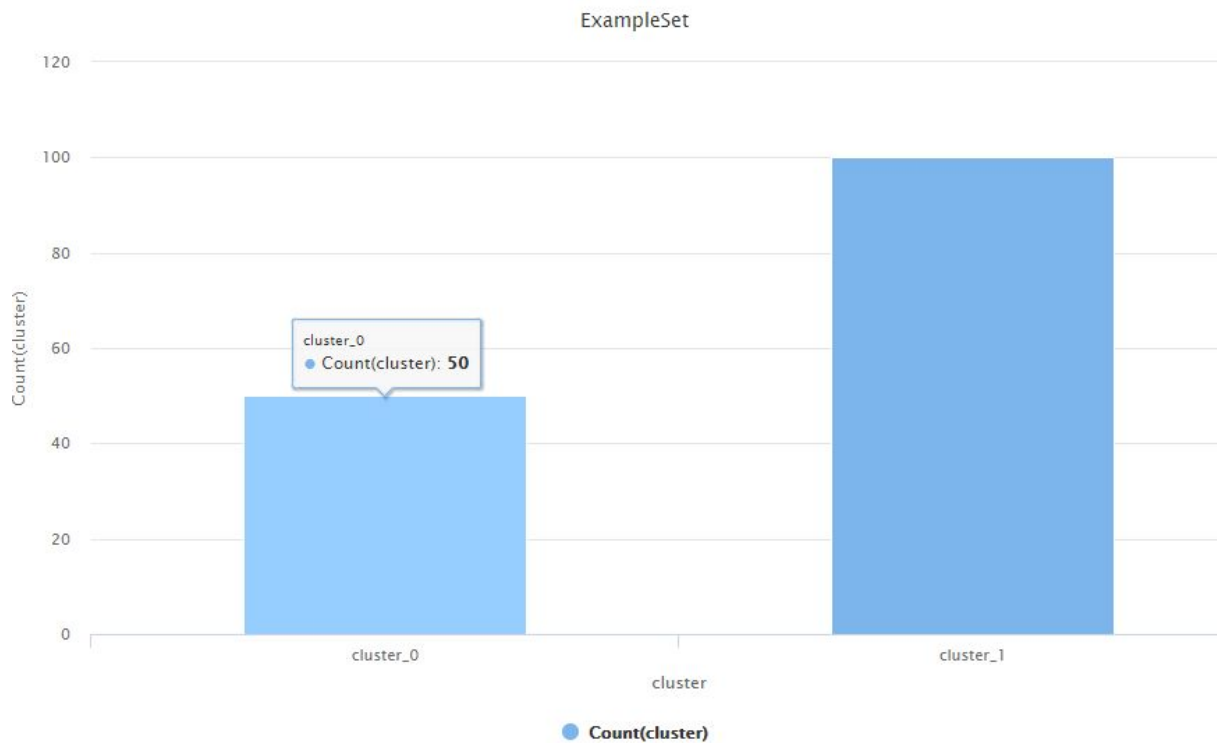
Tareas

1. Prueba de cambiar el número de grupos de agrupamiento y verifique los resultados obtenidos. Considere que el conjunto de datos contiene 3 clases de 50 instancias cada una, donde cada clase se refiere a un tipo de planta de iris. Una clase es linealmente separable de las otras 2; los últimos NO son linealmente separables el uno del otro

A. Identifique los valores de los centroides para diferentes grupos y explique las diferencias

Con K=2 (Centroides)





PerformanceVector (Performance)

Criterion

Avg. within centroid distance

Avg. within centroid distance: 1.482

Criterion

Avg. within centroid distance_cluster_0

Avg. within centroid distance_cluster_0: 0.963

Criterion

Avg. within centroid distance_cluster_1

Avg. within centroid distance_cluster_1: 1.741

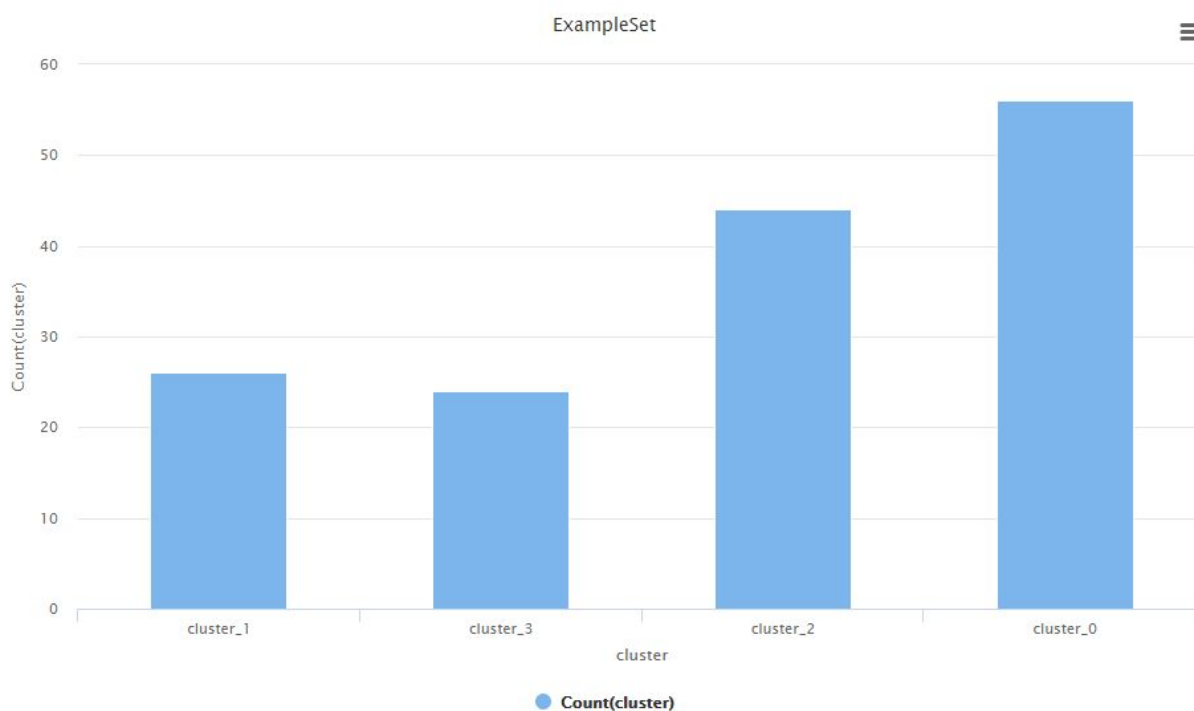
Criterion

Davies Bouldin

Davies Bouldin: 0.598

Con K=4

| ExampleSet (Clustering) | | | | | |
|-------------------------|---------|---------|----------------|----------------|---|
| Name | Type | Missing | Statistics | | Filter (6 / 6 attributes): <input type="text" value="Search for Attributes"/> |
| id | Integer | 0 | Min | Max | Average |
| id | Integer | 0 | 1 | 150 | 75.500 |
| Cluster | Nominal | 0 | Least | Most | Values |
| cluster | Nominal | 0 | cluster_3 (24) | cluster_0 (56) | cluster_0 (56), cluster_2 (44), ...[2 more] |
| longsepal | Real | 0 | Min | Max | Average |
| longsepal | Real | 0 | -1.864 | 2.484 | -0.000 |
| anchosepal | Real | 0 | Min | Max | Average |
| anchosepal | Real | 0 | -2.431 | 3.104 | -0.000 |
| longpetal | Real | 0 | Min | Max | Average |
| longpetal | Real | 0 | -1.563 | 1.780 | -0.000 |
| anchopetal | Real | 0 | Min | Max | Average |
| anchopetal | Real | 0 | -1.440 | 1.705 | -0.000 |



Criterion

Avg. within centroid dis...

Avg. within centroid dis...

Avg. within centroid dis...

Avg. within centroid dis...

Avg. within centroid distance

Avg. within centroid distance: 0.759

Criterion

Avg. within centroid dis...

Avg. within centroid dis...

Avg. within centroid dis...

Avg. within centroid dis...

Avg. within centroid distance_cluster_0

Avg. within centroid distance_cluster_0: 0.867

Criterion

Avg. within centroid dis...

Avg. within centroid dis...

Avg. within centroid dis...

Avg. within centroid distance_cluster_1

Avg. within centroid distance_cluster_1: 0.503

Criterion

Avg. within centroid dis...

Avg. within centroid dis...

Avg. within centroid dis...

Avg. within centroid distance_cluster_2

Avg. within centroid distance_cluster_2: 0.988

Criterion

Avg. within centroid dis...

Avg. within centroid dis...

Avg. within centroid dis...

Avg. within centroid dis...

Avg. within centroid distance_cluster_3

Avg. within centroid distance_cluster_3: 0.363

Criterion

Avg. within centroid dis...

Avg. within centroid dis...


Avg. within centroid dis...

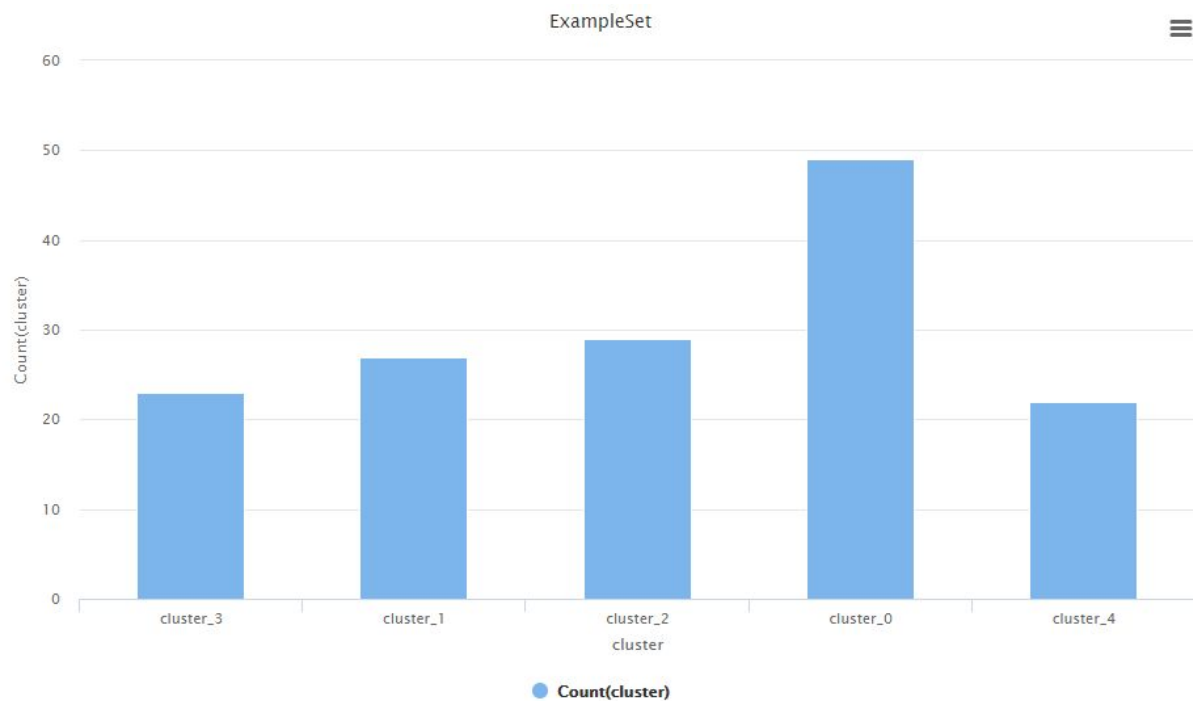
Avg. within centroid dis...

Davies Bouldin

Davies Bouldin: 0.868

Con K=5

| | | | | | |
|----------------------|---------|---|--|------------------------|---|
| ✓ Id | Integer | 0 | Min 1 | Max 150 | Average 75.500 |
| ✓ Cluster | Nominal | 0 | Least cluster_4 (22) | Most cluster_0 (49) | Values cluster_0 (49), cluster_2 (29), ...[3 more] |
| ✓ longsepal | Real | 0 | Min -1.864 | Max 2.484 | Average -0.000 |
| ✓ anchosepal | Real | 0 | Min -2.431 | Max 3.104 | Average -0.000 |
| ✓ longpetal | Real | 0 | Min -1.563 | Max 1.780 | Average -0.000 |
| ✓ anchopetal | Real | 0 | Min -1.440 | Max 1.705 | Average -0.000 |
| Cluster ✓ cluster | Nominal | 0 |  <div>Least cluster_4 (22)</div> <div>Most cluster_0 (49)</div> <div>cluster_0 (49), cluster_1 (27), ...[1 more]</div> <div>Open visualizations</div> <div>Details...</div> | | |



Avg. within centroid distance

Avg. within centroid distance: 0.603

Avg. within centroid distance_cluster_0

Avg. within centroid distance_cluster_0: 0.578

Avg. within centroid distance_cluster_1

Avg. within centroid distance_cluster_1: 0.385

Avg. within centroid distance_cluster_2

Avg. within centroid distance_cluster_2: 0.931

Avg. within centroid distance_cluster_3

Avg. within centroid distance_cluster_3: 0.494

Avg. within centroid distance_cluster_4

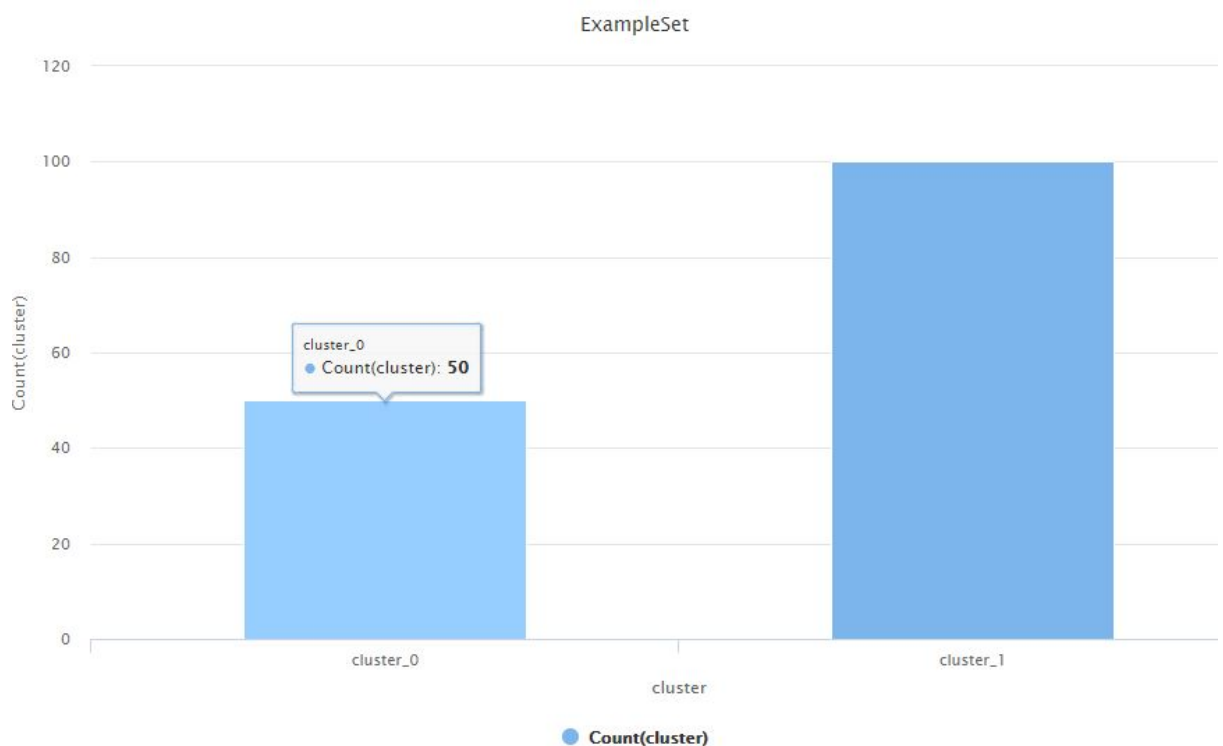
Avg. within centroid distance_cluster_4: 0.611

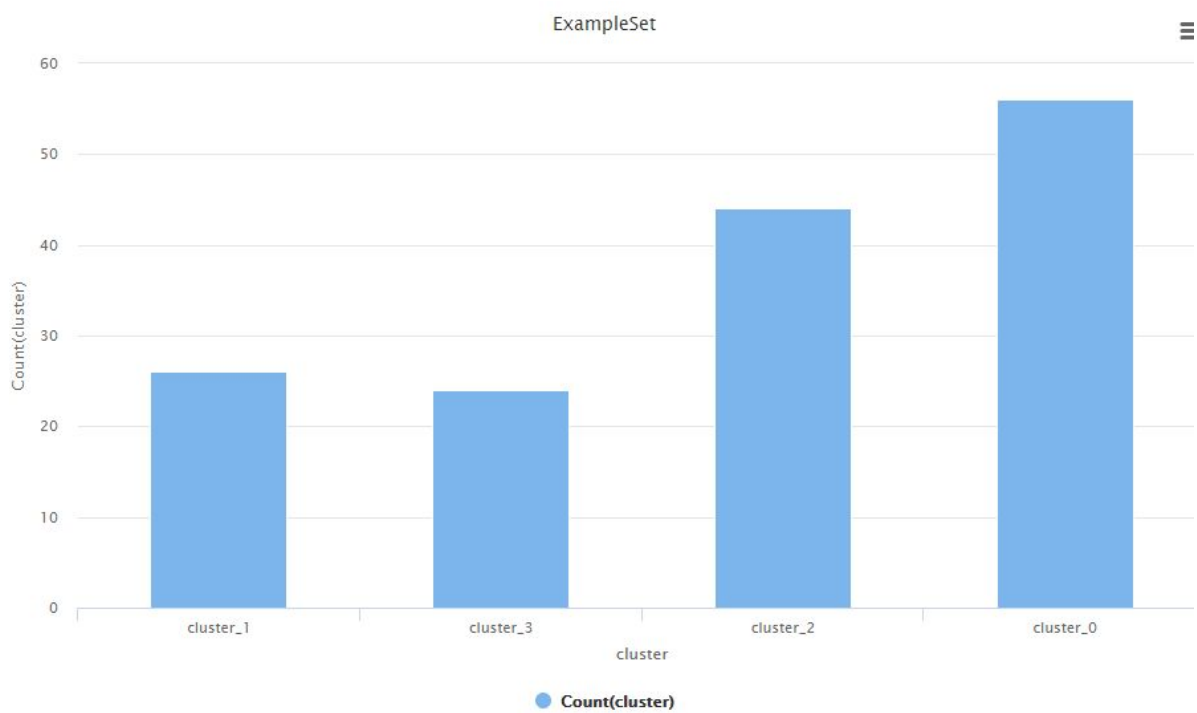
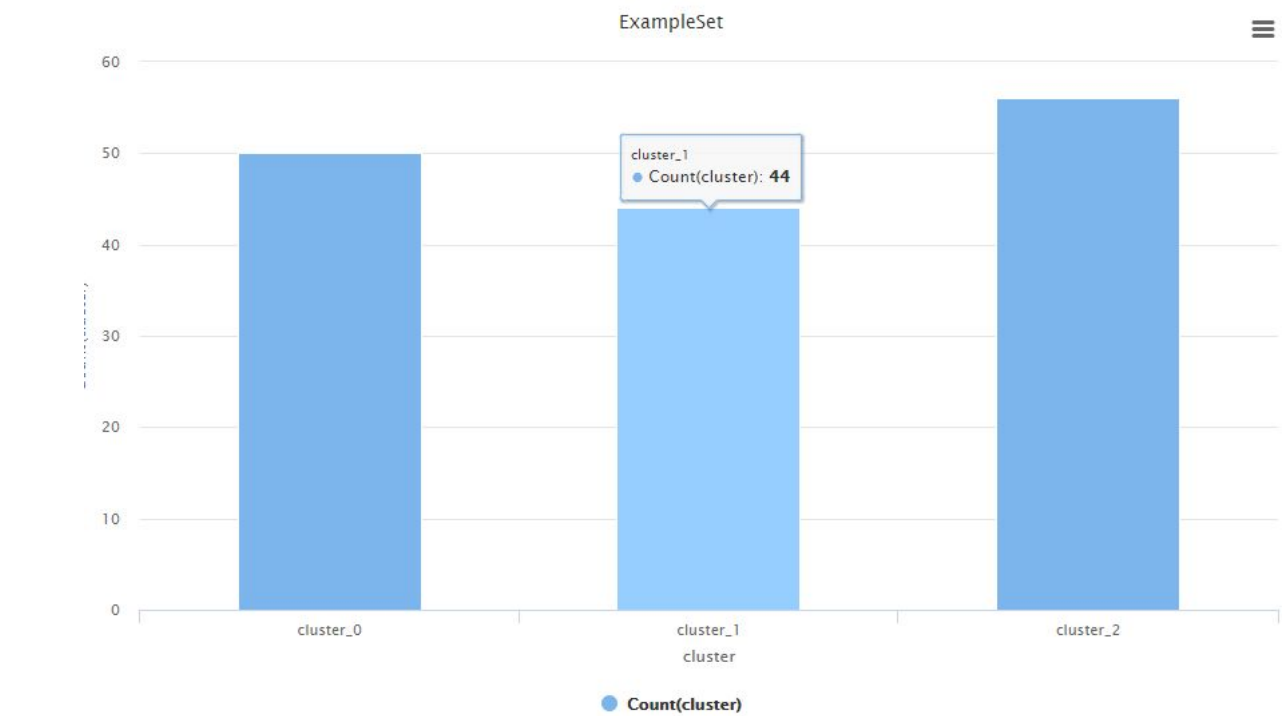
Davies Bouldin

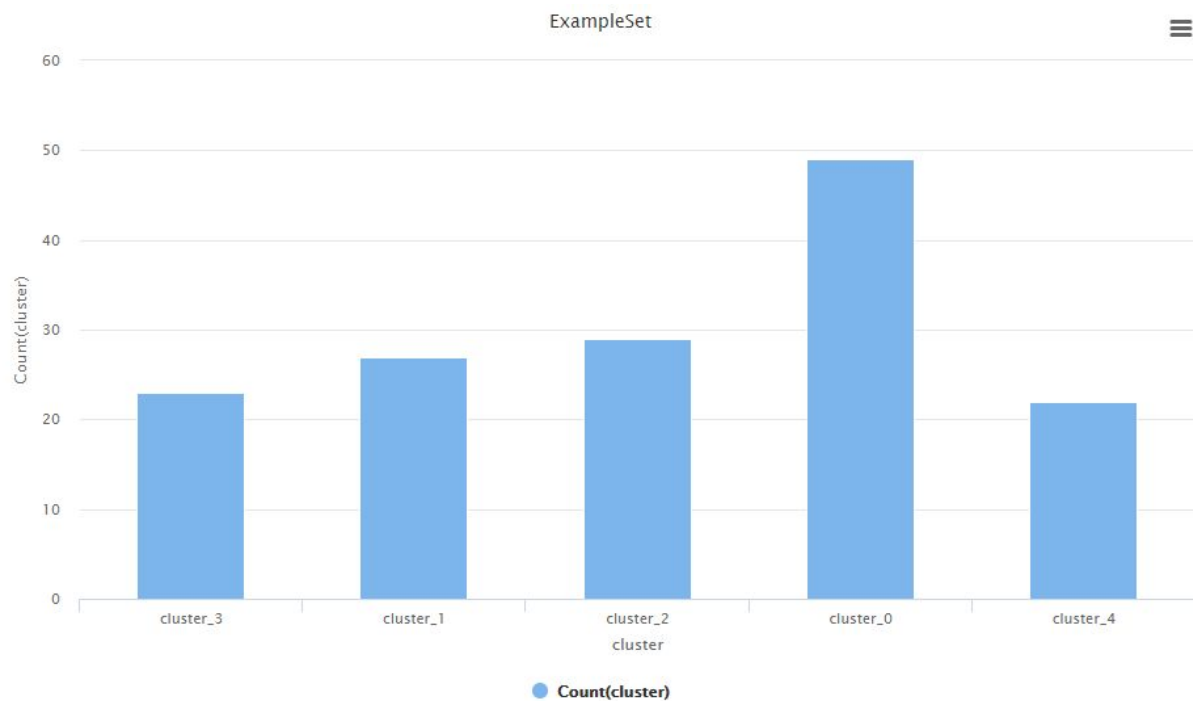
Davies Bouldin: 0.940

Dadas esta comparación, dados diversos números de centroides (2,3,4,5) se encuentra que el cluster con un número de centroides igual a 3, cumple la condicion de tener la máxima distancia entre centroides y la mínima distancia entre puntos con su centroide.

- B. Use los gráficos disponibles usando diferentes parámetros para decidir el mejor número de cluster?







Dadas esta comparacion, dados diversos números de centroides (2,3,4,5) se encuentra que el cluster con un número de centroides igual a 3, cumple la condicion de tener la máxima distancia entre centroides y la mínima distancia entre puntos con su centroide.