

Módulo 4 Metodologías de ciencia de datos

Dr. Carlos Alberto González Martínez



k-Means

Cluster Analysis

gmc

Objetivo

El participante desarrollará la interpretación de los resultados de los modelos trabajados.

METODOLOGÍA DE CIENCIA DE DATOS

EJEMPLO PRÁCTICO

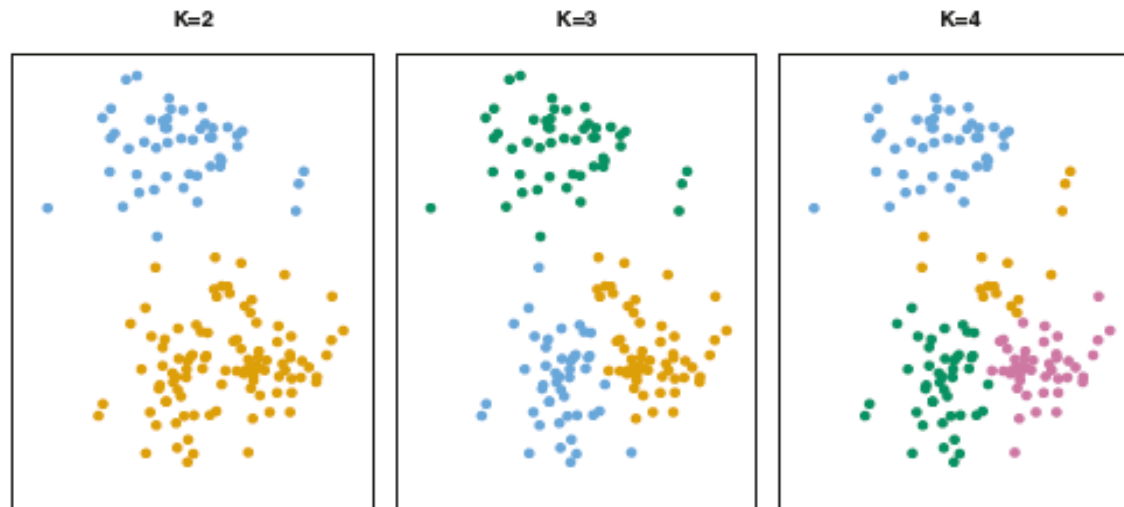
Contenido

Ejemplo práctico en RapidMiner

1. **K-means**
2. Clustering por K-means
3. El algoritmo K-means
4. Clustering
5. Clustering K-means
6. Interpretación de resultados

K-means

El algoritmo K-medias propuesto por [MacQueen en el año 1967](#), es un algoritmo que permite descubrir agrupamientos en conjuntos de datos.



gmc

K-means

K-medias es un método que tiene como objetivo generar una partición de un conjunto de n observaciones en k grupos.

Cada grupo está representado por el promedio de los puntos que lo componen. El representante de cada grupo se denomina centroide. La cantidad de grupos a descubrir, k , es un parámetro que se debe fijar a priori.

El método de clustering comienza con k centroides ubicados de forma aleatoria, y asigna cada observación al centroide más cercano. Después de asignarlos, los centroides se mueven a la ubicación promedio de todos los datos asignados a él, y se vuelven a reasignar los puntos de acuerdo a las nuevas posiciones de los centroides.

K-means

El objetivo de K-medias es agrupar las observaciones de forma tal que todas las que se encuentren en el mismo grupo sean lo más semejantes entre sí y que las pertenecientes a grupos distintos sean lo más desemejantes entre sí.

Las medidas de distancia, como la euclídea, son utilizadas para medir la semejanza y desemejanza.

Una medida para indicar cuán bien los centroides representan a los miembros de su grupo, es la suma de los errores al cuadrado.

K-medias, en cada iteración, intenta reducir el valor de la suma de los errores al cuadrado.

gmc

K-means

La medida consiste en la sumatoria de las distancias al cuadrado de cada observación al centroide de su grupo:

Obtener las asignaciones, S , que minimizan la fórmula

Cantidad de grupos

Centroide del grupo i

$$\arg \min_S \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \mu_i\|^2$$

Por cada punto asignado al grupo i

gmc

K-means

El algoritmo siempre termina, ya que no necesariamente encuentra la configuración más óptima, la que se corresponde con el mínimo de la función objetivo.

Hallar un mínimo de la función, a pesar de que no se trate del mínimo absoluto, garantiza un agrupamiento en el que los grupos son poco dispersos y se encuentran separados entre sí.

El algoritmo es significativamente sensible a los centroides que se seleccionan inicialmente de manera aleatoria. Este efecto se puede reducir realizando varias corridas del método.

METODOLOGÍA DE CIENCIA DE DATOS

EJEMPLO PRÁCTICO

Contenido

Ejemplo práctico en RapidMiner

1. K-means
2. **Clustering por K-means**
3. El algoritmo K-means
4. Clustering
5. Clustering K-means
6. Interpretación de resultados

K-means

Clustering por K-means

El algoritmo de las K-medias (presentado por MacQueen en 1967) es uno de los algoritmos de aprendizaje no supervisado más simples para resolver el problema de la clusterización. El procedimiento aproxima por etapas sucesivas un cierto número (prefijado) de clústeres, haciendo uso de los centroides de los puntos que deben representar.

gmc

METODOLOGÍA DE CIENCIA DE DATOS

EJEMPLO PRÁCTICO

Contenido

Ejemplo práctico en RapidMiner

1. K-means
2. Clustering por K-means
3. **El algoritmo K-means**
4. Clustering
5. Clustering K-means
6. Interpretación de resultados

K-means

El algoritmo K-means se compone de los siguientes pasos:

1. Sitúa K puntos en el espacio en el que "viven" los objetos que se quieren clasificar. Estos puntos representan los centroides iniciales de los grupos.
2. Asigna cada objeto al grupo que tiene el centroide más cercano.
3. Tras haber asignado todos los objetos, recalcula las posiciones de los K centroides.
4. Repite los pasos 2 y 3 hasta que los centroides se mantengan estables. Esto produce una clasificación de los objetos en grupos, que permite dar una métrica entre ellos.

Aunque se puede probar que este algoritmo siempre termina, no siempre la distribución que se alcanza es la óptima, ya que es muy sensible a las condiciones iniciales.

gmc

METODOLOGÍA DE CIENCIA DE DATOS

EJEMPLO PRÁCTICO

Contenido

Ejemplo práctico en RapidMiner

1. K-means
2. Clustering por K-means
3. El algoritmo K-means
4. **Clustering**
5. Clustering K-means
6. Interpretación de resultados

RapidMiner

Clustering

El **Clustering** es una tarea que consiste en agrupar un conjunto de objetos (no etiquetados), en subconjuntos de objetos llamados **Clústeres**. Cada **Clúster** está formado por una colección de objetos que son similares (o se consideran similares) entre sí, pero que son distintos respecto a los objetos de otros Clústeres.

Clúster: Conjunto de objetos que son similares entre sí.

Clustering: Tarea de dividir un conjunto de objetos en subconjuntos de objetos (Clústeres) similares entre sí.

En el campo del ML, el **Clustering** se enmarca dentro del **aprendizaje no supervisado**; es decir, que para esta técnica solo disponemos de un conjunto de datos de entrada, sobre los que debemos obtener información sobre la estructura del dominio de salida, que es una información de la cual no se dispone.

gmc

RapidMiner

Clustering

En muchos casos, no se puede definir ningún atributo objetivo (etiqueta) y los datos deben ser agrupados automáticamente.

Este procedimiento se denomina "Clustering". RapidMiner soporta un amplio rango de esquemas de clustering, que se pueden utilizar de la misma forma que cualquier otro esquema de aprendizaje.

METODOLOGÍA DE CIENCIA DE DATOS

EJEMPLO PRÁCTICO

Contenido

Ejemplo práctico en RapidMiner

1. K-means
2. Clustering por K-means
3. El algoritmo K-means
4. Clustering
5. **Clustering K-means**
6. Interpretación de resultados

RapidMiner

Clustering K-means

En este ejemplo, se carga el muy conocido conjunto de datos Iris (la etiqueta también se carga, pero solo se utiliza para visualización y comparación y no para construir los clústeres).

Uno de los esquemas más simples de clustering, denominado KMeans, se aplica luego a este conjunto de datos. Después se realiza una reducción de dimensionalidad, para que soporte mejor la visualización del conjunto de datos en 2 dimensiones.

K-medias o Kmeans es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos, en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Es un método utilizado en minería de datos.

RapidMiner

Clustering K-means

Proceso:

Se utilizará RapidMiner Studio V 9.9

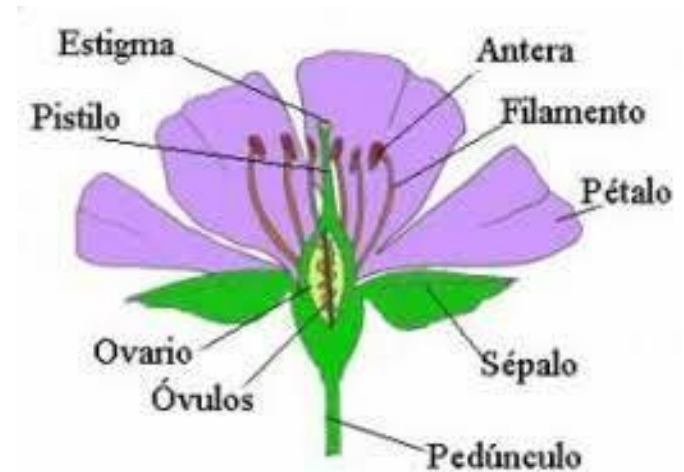
1. Agregar el operador **Data Access** → **Retrieve (Iris)**
2. Agregar el operador **Utility** → **Multiply**
3. Agregar el operador **Modeling** → **Segmentation** → **k-Means**

RapidMiner

Clustering K-means

Base de datos Iris

Contiene: 150 registros
5 columnas o campos:



sepal_length	largo de sépalo
sepal_width	ancho de sépalo
petal_length	largo de pétalo
petal_width	ancho de pétalo
class	especie (setosa, versicolor y virginica)

gmc

RapidMiner

Clustering K-means

Iris setosa



Iris versicolor



Iris virginica



gmc

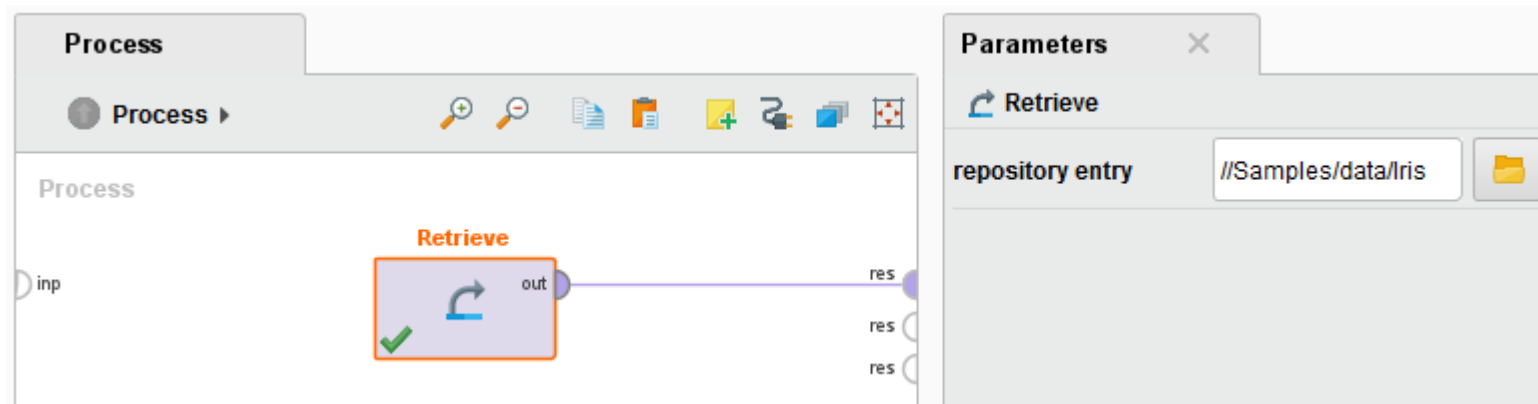
RapidMiner

Clustering K-means

1. Agregar el operador **Data Access** → **Retrieve (Iris)**

Configurarlo para conectarse a los datos de la base de datos de ejemplo Iris.

Conectar la segunda salida **clu** (clustered set) del operador **Retrieve** a un conector **res** del panel.



Ejecutarlo y revisar la salida

gmc


RapidMiner


Clustering K-means

Ejecutarlo y revisar la salida

ExampleSet (Retrieve) ×

Open in

 Turbo Prep

 Auto Model

Filter (150 / 150 examples):

all ▼

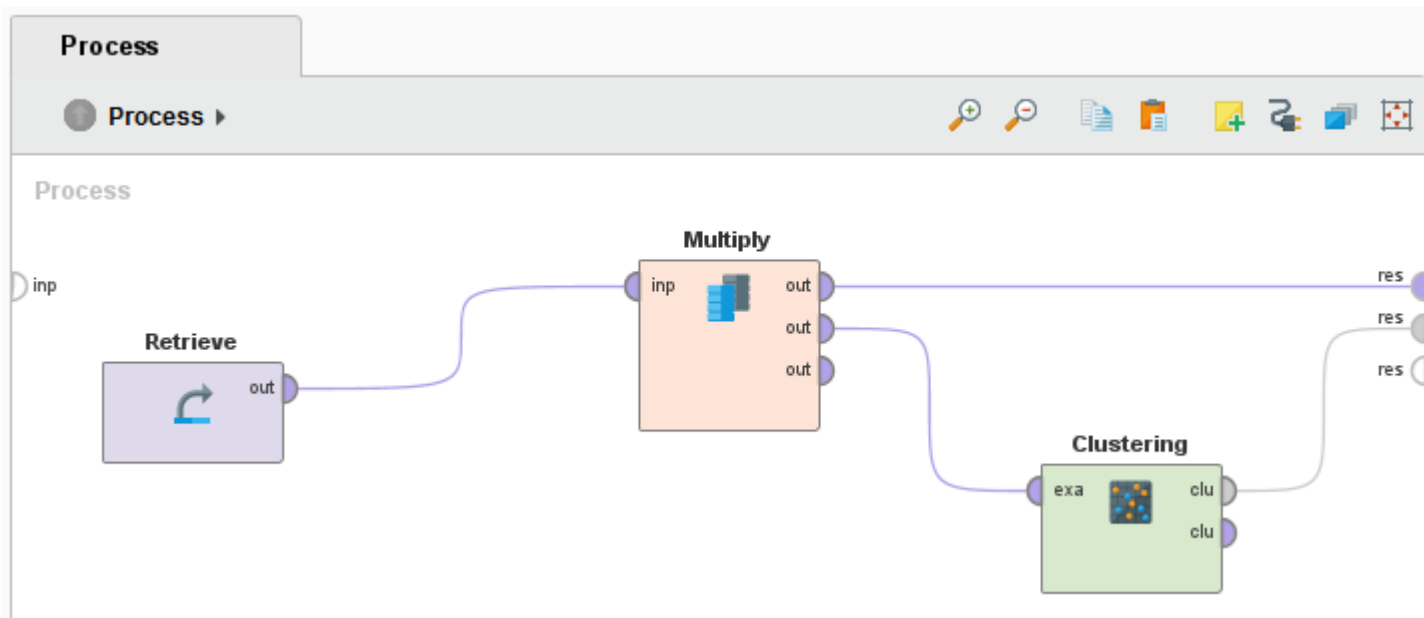
Row No.	id	label	a1	a2	a3	a4
1	id_1	Iris-setosa	5.100	3.500	1.400	0.200
2	id_2	Iris-setosa	4.900	3	1.400	0.200
3	id_3	Iris-setosa	4.700	3.200	1.300	0.200
4	id_4	Iris-setosa	4.600	3.100	1.500	0.200
5	id_5	Iris-setosa	5	3.600	1.400	0.200
6	id_6	Iris-setosa	5.400	3.900	1.700	0.400
7	id_7	Iris-setosa	4.600	3.400	1.400	0.300
8	id_8	Iris-setosa	5	3.400	1.500	0.200
9	id_9	Iris-setosa	4.400	2.900	1.400	0.200
10	id_10	Iris-setosa	4.900	3.100	1.500	0.100
11	id_11	Iris-setosa	5.400	3.700	1.500	0.200

gmc

RapidMiner

Clustering K-means

2. Agregar el operador **Utility** → **Multiply**
3. Agregar el operador **Modeling** → **Segmentation** → **k-Means**



Conectarlo como se muestra en la imagen.


gmc

RapidMiner

Clustering K-means

- Configurar el nodo k-Means (clustering) a 5 clústeres


Parameters

 Clustering (k-Means)

☒ add cluster attribute

☒ add as label

☐ remove unlabeled

k 

5

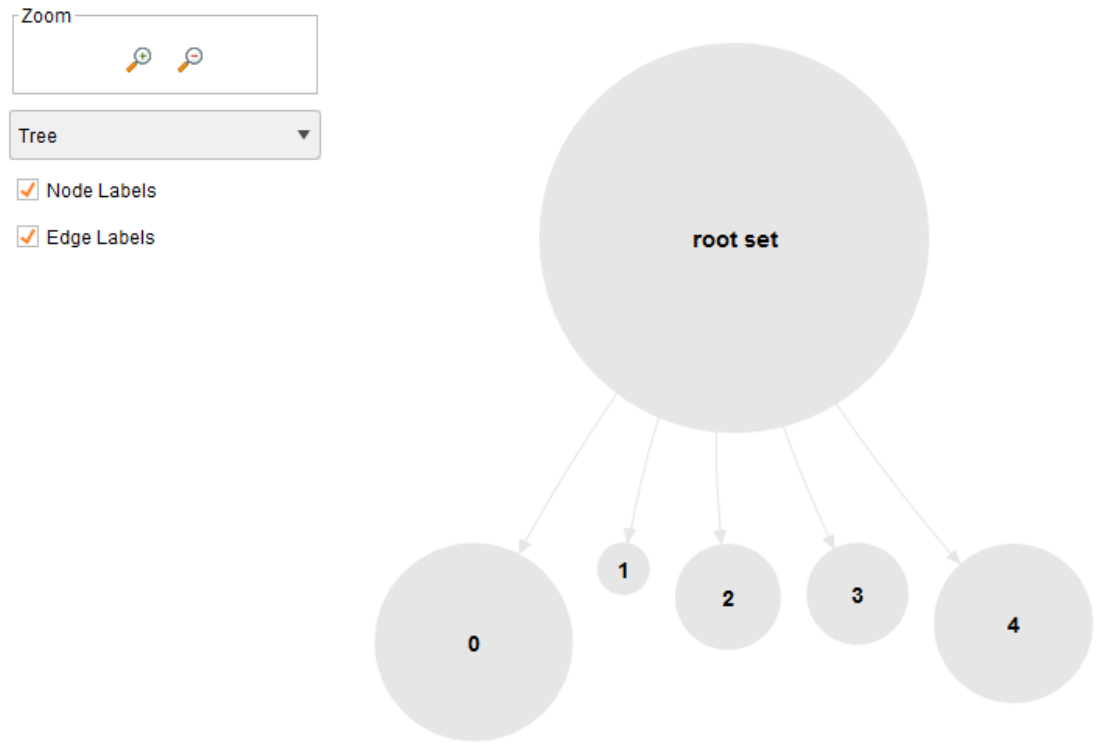
max runs

10

RapidMiner

Clustering K-means

Ejecutarlo y revisar la salida



METODOLOGÍA DE CIENCIA DE DATOS

EJEMPLO PRÁCTICO

Contenido

Ejemplo práctico en RapidMiner

1. K-means
2. Clustering por K-means
3. El algoritmo K-means
4. Clustering
5. Clustering K-means
6. **Interpretación de resultados**

RapidMiner

Interpretación de resultados

Analizar los clústeres generados

Cluster Model

```
Cluster 0: 50 items  
Cluster 1: 12 items  
Cluster 2: 25 items  
Cluster 3: 24 items  
Cluster 4: 39 items  
Total number of items: 150
```

gmc

RapidMiner

Interpretación de resultados

Analizar l< t<bl< de centroides

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
a1	5.006	7.475	5.508	6.529	6.208
a2	3.418	3.125	2.600	3.058	2.854
a3	1.464	6.300	3.908	5.508	4.746
a4	0.244	2.050	1.204	2.162	1.564

Referencias bibliográficas

RapidMiner (2014). RapidMiner Studio Manual, recuperado de <https://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf>

Contacto

Carlos Alberto González Martínez

Jefe de departamento de correlaciones, cruces y alertas (C5i)

gmcmxiv@hotmail.com