Clúster jerárquico con datos: categóricos y nominales

Alumno:
Castelán Contreras Ana Yuleni
De Luna Ocampo Yanina

Profesor:
Fabiola Ocampo Botello

Contenido

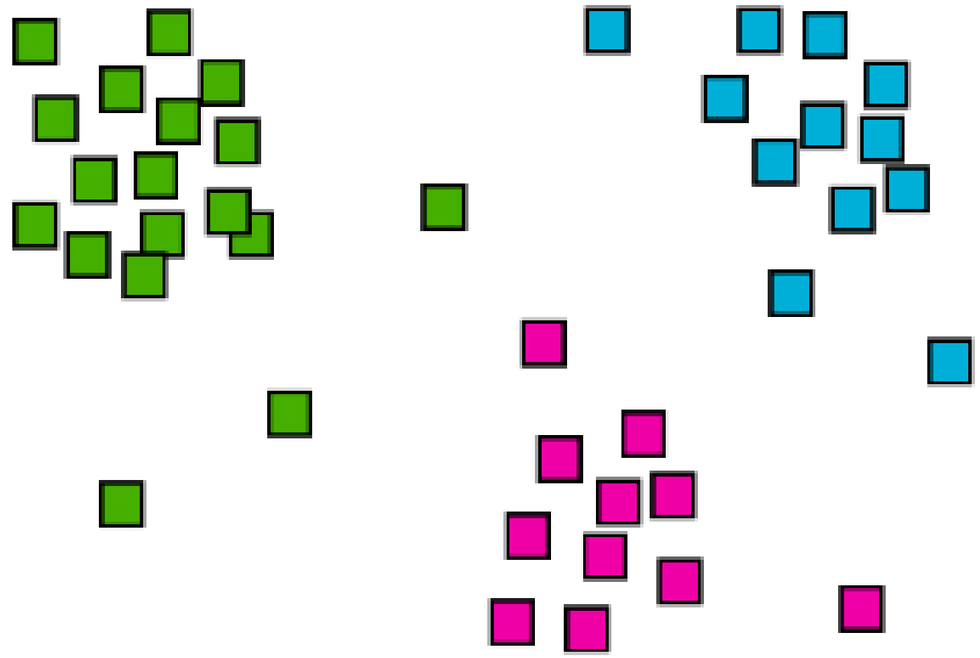
01

¿Qué es un clúster?

Clúster

Es un técnica de minería de datos que agrupa tipos similares de datos o consultas que ayudan a identificar áreas temáticas similares.

La mayoría de estos algoritmos utilizan medidas de distancia para calcular la similitud entre los puntos de datos.



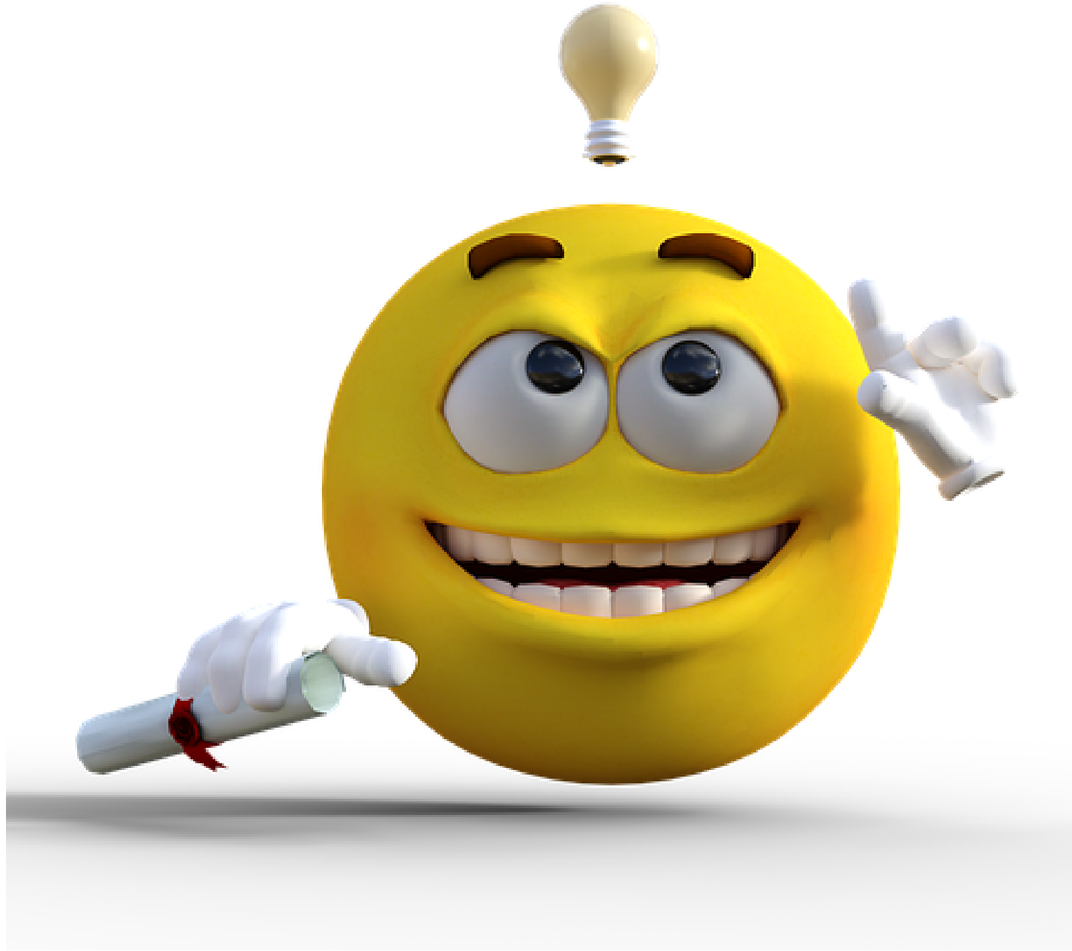
02

ROCK algorithm

RObust Clustering using linKs

Sirve para conjunto de datos con atributos categóricos y booleanos.

Un par de puntos se consideran vecinos si su similitud está por encima del umbral preestablecido.



Caracteristicas

1. Pertenece a la clase de algoritmos jerárquicos aglomerativos que trabajan con datos de tipo categórico
2. Utiliza una muestra aleatoria del conjunto de total de datos para trabajar.
3. Maneja atributos con valores faltantes.
4. Utiliza una funci´on para medir la calidad de la mezcla de los cl´usters.
5. Utiliza ligas para medir el n´umero de vecinos en com´un entre 2 puntos.
6. Mezcla los cl´usters hasta que se obtengan los k cl´usters especificados por el usuario.
7. Descarta del proceso de clustering los datos con pocos o ning´un vecino (outliers).
8. Encuentra cl´usters de formas arbitrarias.

En la vida real es muy común que se presente el problema de analizar datasets con tipos de datos mixtos.

- El algoritmo ROCK evalúa las distancias entre objetos utilizando el coeficiente Jaccard.
- Utiliza el parámetro θ para determinar quienes son los vecinos en cada uno de los objetos.
- Dado un punto p, un punto q es vecino de p si el coeficiente de Jaccard sim(p, q) excede el valor de θ.
- Se generan los valores de la matriz de ligas (links), la cual consiste en la evaluaci´on de links(p, q) como el número de vecinos comunes entre los puntos p y q.

03

Coeficiente de Jaccard

Definición

El algoritmo de similitud Jaccard es un tipo muy interesante de algoritmo aplicado al análisis de grafos.

Esta especialmente diseñado para calcular coeficientes de similitud tomando como base fundamental el método acuñado por Paul Jaccard.

Formula

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Ejemplo 1

Contamos con el siguiente conjunto de datos:
[1, 2, 3] , [1, 2, 4, 5]
Y se busca la similaridad entre los dos calculado con el coeficiente de jaccard

- Comenzamos con su interseccion que en este caso son 2 elementos el 1 y el 2
- Continuamos con el primer conjunto que cuenta con 3 elementos
- Después el segundo conjunto cuenta con 4 elementos
- Aplicando la formula del índice de jaccard obtenemos:

$$2 / 3 + 4 - 2 = 2 / 5 = 0.4$$

Ejemplo 2

Se hace una encuesta a dos personas con 6 preguntas y sus resultados son los siguientes:

	1	2	3	4	5	6
P1	Y	Y	Y	N	N	Y
P2	Y	Y	N	N	Y	Y

Para una mejor lectura pasaremos las respuestas Y a 1 y las respuestas N a 0

	1	2	3	4	5	6
P1	1	1	1	0	0	1
P2	1	1	0	0	1	1

Ejemplo 2

Observamos las similitudes y son las siguientes:

	1	2	3	4	5	6
P1	1	1	1	0	0	1
P2	1	1	0	0	1	1

En este caso no se toman en cuenta los 0 con 0 como intersección debido a que estamos buscando las similitudes de una encuesta en la que se busca saber en que preguntas SI son afines las personas

Así que obtenemos los datos:

$|A \cap B| = 3$

$|A| = 4$

$|B| = 4$

Debido a que solo se toman en cuenta las preguntas respondidas con SI

Ejemplo 2

Y aplicando la formula se obtiene:

$$|A \cap B| \; / \; |A| + |B| - |A \cap B|$$

$$= 3 \; / \; 4 + 4 - 3$$

$$= 3 \; / \; 5$$

$$= 0.6$$

Ejemplo 3

Contamos con el siguiente conjunto de datos:
[Algodón, Cubrebocas, Lampara]
[Algodón, Apósito, Guantes, Hilo, Lampara]
Y se busca la similaridad entre los dos calculado con el coeficiente de jaccard

- Comenzamos con su intersección que en este caso son 3 elementos
- Continuamos con el primer conjunto que cuenta con 3 elementos
- Después el segundo conjunto cuenta con 5 elementos
- Aplicando la formula del índice de jaccard obtenemos:

$$3 / 3 + 5 - 3 = 3 / 5 = 0.6$$

Ejemplo 3

Otra forma de representar el anterior conjunto de datos son:

	Algohodon	Aposito	Cubrebocas	Guantes	Hilo	Lampara
P1	1	0	1	0	0	1
P2	1	1	0	1	1	1

Al representarse de esta manera podemos observar porque no se toman en cuenta los elementos con 0 para el conjunto |A| o |B| ya que el conjunto original no cuenta con estos y si hubiera casos con relación 00 no contaria debido a que ese conjunto a pesar de establecerlo no esta siendo tomado en cuenta

04

Valor umbral θ

Definición

El funcionamiento de ROCK está basado en el concepto de los enlaces (links) entre objetos vecinos (neighbors).

Se describe como vecinos a dos objetos si el valor de una función de similitud entre los dos objetos excede cierto valor de límite θ.

$sim(\hat{x}, \hat{y}) \geq \theta$

El valor umbral θ está definido por el usuario.

Es decir si la similitud entre dos puntos (dada por el indice de jaccard) es menor o igual al valor umbral θ esto quiere decir que esos dos puntos son vecinos

05

Ejemplo

	A	B	C
1	1	0	1
2	0	1	1
3	1	1	1
4	1	1	0

	$\lvert A \cap B \rvert$	$\lvert A \rvert$	$\lvert B \rvert$	CJ		$\lvert A \cap B \rvert$	$\lvert A \rvert$	$\lvert B \rvert$	CJ
1,2	1	2	2	1/3	2,3	2	2	3	2/3
1,3	2	2	3	2/3	2,4	1	2	3	1/4
1,4	1	2	3	1/4	3,4	2	3	3	1/2

	1	2	3	4
1	1	0	0	0
2	0.33	1	0	0
3	0.66	0.66	1	0
4	0.25	0.25	0.50	1

	1	2	3	4
1	1	0	0	0
2	0.33	1	0	0
3	0.66	0.66	1	0
4	0.25	0.25	0.50	1

	1	2	3	4
1	1	0	0	0
2	1	1	0	0
3	1	1	1	0
4	0	0	1	1

	1	2	3	4
1	1	0	0	0
2	1	1	0	0
3	1	1	1	0
4	0	0	1	1

	1	2	3	4
1	1	0	0	0
2	2	1	0	0
3	3	2	1	0
4	1	1	2	1

06

Ejercicio en clase

	A	B	C	D	E	F	G
1	1	0	0	0	1	1	0
2	0	0	1	1	0	1	0
3	1	1	1	1	0	0	1
4	0	0	0	1	1	0	1
5	0	1	0	1	0	1	1
6	0	1	1	0	0	1	1
7	1	1	0	1	0	1	1
8	1	1	1	1	0	0	0
9	1	1	0	1	0	0	1
10	0	1	1	0	0	0	0

	$\lvert A \cap B \rvert$	$\lvert A \rvert$	$\lvert B \rvert$	CJ		$\lvert A \cap B \rvert$	$\lvert A \rvert$	$\lvert B \rvert$	CJ		$\lvert A \cap B \rvert$	$\lvert A \rvert$	$\lvert B \rvert$	CJ
1,2					2,4					3,7				
1,3					2,5					3,8				
1,4					2,6					3,9				
1,5					2,7					3,10				
1,6					2,8					4,5				
1,7					2,9					4,6				
1,8					2,10					4,7				
1,9					3,4					4,8				
1,10					3,5					4,9				
2,3					3,6					4,10				

	$\lvert A \cap B \rvert$	$\lvert A \rvert$	$\lvert B \rvert$	IJ		$\lvert A \cap B \rvert$	$\lvert A \rvert$	$\lvert B \rvert$	IJ		$\lvert A \cap B \rvert$	$\lvert A \rvert$	$\lvert B \rvert$	IJ
1,2	1	3	3	1/5	2,4	1	3	3	1/5	3,7	4	5	5	2/3
1,3	1	3	5	1/7	2,5	2	3	4	2/5	3,8	4	5	4	4/5
1,4	1	3	3	1/5	2,6	2	3	4	2/5	3,9	4	5	4	4/5
1,5	1	3	4	1/6	2,7	2	3	5	1/3	3,10	2	5	2	2/5
1,6	1	3	4	1/6	2,8	2	3	4	2/5	4,5	2	3	4	2/5
1,7	2	3	5	1/3	2,9	1	3	4	1/6	4,6	1	3	4	1/6
1,8	1	3	4	1/6	2,10	1	3	2	1/4	4,7	2	3	5	1/3
1,9	1	3	4	1/6	3,4	2	5	3	1/3	4,8	1	3	4	1/6
1,10	0	3	2	0	3,5	3	5	4	1/2	4,9	2	3	4	2/5
2,3	2	3	5	1/3	3,6	3	5	4	1/2	4,10	0	3	2	0

	|A ∩ B|	|A|	|B|	IJ		|A ∩ B|	|A|	|B|	IJ
5,6					6,10				
5,7					7,8				
5,8					7,9				
5,9					7,10				
5,10					8,9				
6,7					8,10				
6,8					9,10				
6,9					-	-	-	-	-

	$\lvert A \cap B \rvert$	$\lvert A \rvert$	$\lvert B \rvert$	IJ		$\lvert A \cap B \rvert$	$\lvert A \rvert$	$\lvert B \rvert$	IJ
5,6	3	4	4	3/5	6,10	2	4	2	1/2
5,7	4	4	5	4/5	7,8	3	5	4	1/2
5,8	2	4	4	1/3	7,9	4	5	4	4/5
5,9	3	4	4	3/5	7,10	1	5	2	1/6
5,10	1	4	2	1/5	8,9	3	4	4	3/5
6,7	3	4	5	1/2	8,10	2	4	2	1/2
6,8	2	4	4	1/3	9,10	1	4	2	1/5
6,9	2	4	4	1/3	-	-	-	-	-

	1	2	3	4	5	6	7	8	9	10
1	1	-	-	-	-	-	-	-	-	-
2		1	-	-	-	-	-	-	-	-
3			1	-	-	-	-	-	-	-
4				1	-	-	-	-	-	-
5					1	-	-	-	-	-
6						1	-	-	-	-
7							1	-	-	-
8								1	-	-
9									1	-
10										1

	1	2	3	4	5	6	7	8	9	10
1	1	-	-	-	-	-	-	-	-	-
2	0.20	1	-	-	-	-	-	-	-	-
3	0.14	0.33	1	-	-	-	-	-	-	-
4	0.20	0.20	0.33	1	-	-	-	-	-	-
5	0.16	0.40	0.50	0.40	1	-	-	-	-	-
6	0.16	0.40	0.50	0.16	0.60	1	-	-	-	-
7	0.33	0.33	0.66	0.33	0.80	0.50	1	-	-	-
8	0.16	0.40	0.80	0.16	0.33	0.33	0.50	1	-	-
9	0.16	0.16	0.80	0.40	0.60	0.33	0.80	0.60	1	-
10	0	0.25	0.40	0	0.20	0.50	0.16	0.50	0.20	1

	1	2	3	4	5	6	7	8	9	10
1	1	-	-	-	-	-	-	-	-	-
2	0.20	1	-	-	-	-	-	-	-	-
3	0.14	0.33	1	-	-	-	-	-	-	-
4	0.20	0.20	0.33	1	-	-	-	-	-	-
5	0.16	0.40	0.50	0.40	1	-	-	-	-	-
6	0.16	0.40	0.50	0.16	0.60	1	-	-	-	-
7	0.33	0.33	0.66	0.33	0.80	0.50	1	-	-	-
8	0.16	0.40	0.80	0.16	0.33	0.33	0.50	1	-	-
9	0.16	0.16	0.80	0.40	0.60	0.33	0.80	0.60	1	-
10	0	0.25	0.40	0	0.20	0.50	0.16	0.50	0.20	1

	1	2	3	4	5	6	7	8	9	10
1	1	-	-	-	-	-	-	-	-	-
2	0	1	-	-	-	-	-	-	-	-
3	0	0	1	-	-	-	-	-	-	-
4	0	0	0	1	-	-	-	-	-	-
5	0	0	0	0	1	-	-	-	-	-
6	0	0	0	0	1	1	-	-	-	-
7	0	0	1	0	1	0	1	-	-	-
8	0	0	1	0	0	0	0	1	-	-
9	0	0	1	0	1	0	1	1	1	-
10	0	0	0	0	0	0	0	0	0	1

	1	2	3	4	5	6	7	8	9	10
1	1	-	-	-	-	-	-	-	-	-
2	0	1	-	-	-	-	-	-	-	-
3	0	0	1	-	-	-	-	-	-	-
4	0	0	0	1	-	-	-	-	-	-
5	0	0	0	0	1	-	-	-	-	-
6	0	0	0	0	2	1	-	-	-	-
7	0	0	2	0	2	0	1	-	-	-
8	0	0	2	0	0	0	0	1	-	-
9	0	0	4	0	3	0	2	2	1	-
10	0	0	0	0	0	0	0	0	0	1