ANALISIS DE CLUSTER

Dr. Porfirio Gutiérrez González

Análisis de Clúster

El análisis de clúster es una técnica de análisis de la interdependencia cuyo fin es clasificar objetos o sujetos en función de las variables: la idea es formar grupos con respecto a las variables, donde las diferencias entre los sujetos pertenecientes a un mismo grupo sean mínimas y las diferencias respecto a sujetos de otros grupos sea máxima.

Para realizar un análisis de clúster se necesita:

- **Obtención de la matriz de datos.**
- **Estandarización de la matriz de datos (opcional)**
- **Calculo de la matriz de semejanzas o distancias.**
- **Ejecución del método de agrupamiento.**

Obtención de la matriz de datos.

Para un análisis de clúster se requiere de una matriz de datos X, donde las filas corresponden a n-sujetos y las columnas corresponden a p-variables medidas.

Las p-variables medidas que describen a cada sujeto, constituyen un aspecto importante en la formación de los grupos de los sujetos, por consiguiente, el investigador debe seleccionar las variables adecuadamente y las variables que identifican a los sujetos.

Estandarización de la matriz de datos.

En una matriz de datos X, se puede encontrar que algunas variables estén medidas en diferentes unidades, por ejemplo, una en kilogramos, otra en centímetros, otra en pulgadas, otra en volumen: para esto se requiere estandarizar las variables, para que las diferencias de unidades de medida no afecten a los resultados y de esta forma cada variable entre al análisis en igualdad de circunstancias.

Métrica de distancia o similitud

Las métricas de distancia o de similitud son utilizadas para medir la semejanza o proximidad de 2 objetos o individuos con respecto a sus características, generalmente conocidas como variables.

Métrica Euclidiana

- La medida más usada comúnmente para cuantificar el grado de similitud entre dos objetos u individuos es conocida como la métrica Euclidiana.

- **La métrica Euclidiana mide la distancia entre dos sujetos cuando son vistos como puntos de un espacio p-dimensional formado por sus variables.**

$$\delta_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ij} - x_{jk}\right)^2}$$

Otra distancia que también se utiliza en este tipo de datos, es la distancia de Minkowsky

$$\boldsymbol{d_r(i,j) = \left(\sum_{k=1}^{p} \left|x_{ik} - x_{jk}\right|^{r}\right)^{1/r}}$$

Cuando r=1se le denomina distancia media o bloque de ciudad.

$$\boldsymbol{d_1(i,j) = \left(\sum_{k=1}^{p} \left|x_{ik} - x_{jk}\right|\right)}$$

$$\delta_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ij} - x_{jk})^2}$$

EJEMPLO

individuo	X1=ESTATURA	X2=PESO
1	1.50	62
2	1.60	63
3	1.65	75
4	1.70	80
5	1.85	95
6	1.90	100

$$\delta_{12} = \sqrt{(1.5 - 1.60)^2 + (62 - 63)^2} = 1$$

$$\delta_{13} = \sqrt{(1.5 - 1.65)^2 + (62 - 75)^2} = 13$$

$$\delta_{14} = \sqrt{(1.5 - 1.70)^2 + (62 - 80)^2} = 18$$

$$\delta_{15} = \sqrt{(1.5 - 1.85)^2 + (62 - 95)^2} = 33$$

$$\delta_{16} = \sqrt{(1.5 - 1.90)^2 + (62 - 100)^2} = 38$$

$$\delta_{23} = \sqrt{(1.60 - 1.65)^2 + (63 - 75)^2} = 12$$

$$\delta_{24} = \sqrt{(1.6 - 1.7)^2 + (63 - 80)^2} = 17$$

$$\delta_{25} = \sqrt{(1.6 - 1.85)^2 + (63 - 95)^2} = 32$$

$$\delta_{26} = \sqrt{(1.6 - 1.9)^2 + (63 - 100)^2} = 37$$

$$\delta_{34} = \sqrt{(1.65 - 1.70)^2 + (75 - 80)^2} = 5$$

$$\delta_{35} = \sqrt{(1.65 - 1.85)^2 + (75 - 95)^2} = 20$$

$$\delta_{36} = \sqrt{(1.65 - 1.9)^2 + (75 - 100)^2} = 25$$

$$\delta_{45} = \sqrt{(1.70 - 1.85)^2 + (80 - 95)^2} = 15$$

$$\delta_{46} = \sqrt{(1.70 - 1.9)^2 + (80 - 100)^2} = 20$$

$$\delta_{56} = \sqrt{(1.85 - 1.9)^2 + (90 - 100)^2} = 5$$

MATRIZ DE DISTANCIA O SIMILITUD

	1	2	3	4	5	6
1		1	13	18	33	38
2	1		12	17	32	37
3	13	12		5	20	25
4	18	17	5		15	20
5	33	32	20	15		5
6	38	37	25	20	5	

- **Vecino más cercano.**

Define d (i, j) como la disimilaridad más pequeña entre una entidad R y una entidad de P. Este método no resulta apropiado en aquellas situaciones en la que los dos grupos se acercan demasiado, ya que inmediatamente pasarían a estar unidos (y ya no podrían separarse en los siguientes pasos). Este fenómeno se conoce con el nombre de encadenamiento y da como resultado conglomerados alargados, en los que algunos miembros se encuentran muy alejados de otros.

- **Vecino más lejano.**

Lo opuesto a lo anterior: utiliza la disimilaridad más grande entre una entidad de R y una de ς. Lógicamente, es entonces la técnica menos posible a producir encadenamiento. Produce conglomerados compactos, de pequeño diámetro. Pero no tienen por qué estar finalmente bien separados entre sí, ya que tienden a no agruparse cuando contienen elementos demasiados distantes.

- **Método de Ward**

También llamado método de la varianza mínima, ya que lo que hace es buscar dos conglomerados cuya unión con lleve el menor incremento de la varianza. Esto significa que en cada paso debe probar todas las combinaciones posibles de dos grupos, calcular el valor del índice de la suma de cuadrados y seleccionar aquel de menor valor.

- **Ejemplo:Vecino más cercano.**

	1	2	3	4	5	6
1		1	13	18	33	38
2	1		12	17	32	37
3	13	12		5	20	25
4	18	17	5		15	20
5	33	32	20	15		5
6	38	37	25	20	5	

$minimo\ (\delta_{13,}\ \delta_{23}) = (13,12) = 12$

$minimo\ (\delta_{14,}\ \delta_{24}) = (18,17) = 17$

$minimo\ (\delta_{15,}\ \delta_{25}) = (33,32) = 32$

$minimo\ (\delta_{16,}\ \delta_{26}) = (38,37) = 37$

	1,2	3	4	5	6
1,2		13	18	32	37
3	12		5	20	25
4	17	5		15	20
5	32	20	15		5
6	37	25	20	5	

$minimo\ (\delta_{13,}\ \delta_{14},\ \delta_{23},\ \delta_{24}) = (13,18,12,17) = 12$

$minimo\ (\delta_{35,}\ \delta_{45}) = (20,15) = 15$

$minimo\ (\delta_{36,}\ \delta_{46}) = (25,20) = 20$

	1,2	3,4	5	6
1,2		12	32	37
3,4	12		15	25
5	32	15		5
6	37	20	5	

- **Ejemplo:Vecino más cercano.**

	1,2	3,4	5,6
1,2		12	32
3,4	12		15
5,6	32	15	

$$minimo\ (\delta_{15,}\ \delta_{16},\ \delta_{25},\ \delta_{26}) = (33,38,32,37) = 32$$

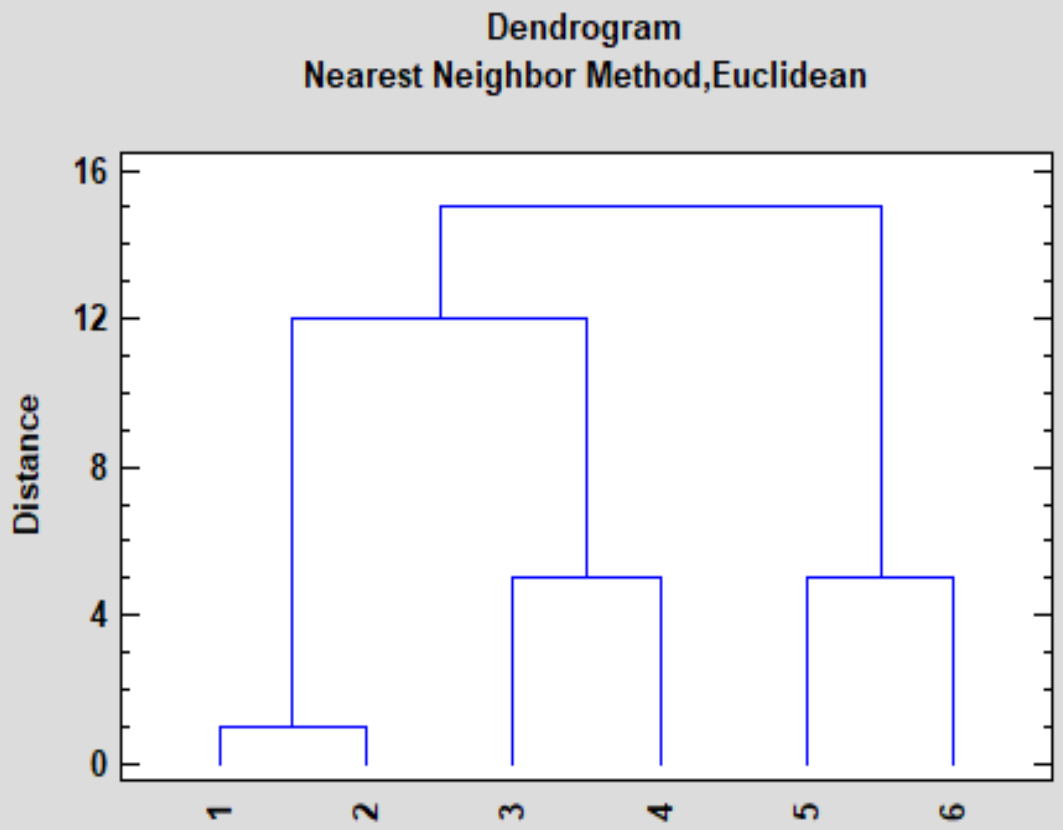
$$minimo\ (\delta_{35,}\ \delta_{36},\ \delta_{45},\ \delta_{46}) = (20,25,15,20) = 15$$

	1,2,3,4	5,6
1,2,3,4		15
5,6	15	

	1	2	3	4	5	6
1		1	13	18	33	38
2	1		12	17	32	37
3	13	12		5	20	25
4	18	17	5		15	20
5	33	32	20	15		5
6	38	37	25	20	5	

$$minimo\ (\delta_{15,}\ \delta_{16},\ \delta_{25},\ \delta_{26},\ \delta_{35},\ \delta_{36},\ \delta_{45},\ \delta_{46}) = (33,38,32,37,20,25,15,20) = 15$$

DENDOGRAMA


Dendrogram
Nearest Neighbor Method,Euclidean
Distance
16
12
8
4
0
1
2
3
4
5
6