UNIVERSITAT DE BARCELONA

Cluster analysis

Píndoles d'estadística avançada
STeL (Març 2021)
Sessió 1

PROF. S. CIVIT

1

Cluster analysis

- L´anàlisi de grups (cluster analysis) és un conjunt de tècniques estadístiques que serveixen per **determinar grups internament homogenis, però diferents entre ells**, bé per agrupació d´unitats més petites (bottom-up), o per divisió de segments majors (top-down)
- Aquestes tècniques agrupen tant individus com variables

2

Hierarchical vs Non Hierarchical

- Hierarchical clustering: Agglomerative vs Divisive
- Existeix una gran varietat de tècniques d'anàlisis de grups. Tenim els ”ascendents” (bottom-up) que construeixen els grups per agregació (**agglomerative**), a partir dels individus considerats un a un, i els “descendents” (top-down)”, parteixen del conjunt total d'individus i els divideixen (**divisive)** en grups més petits.

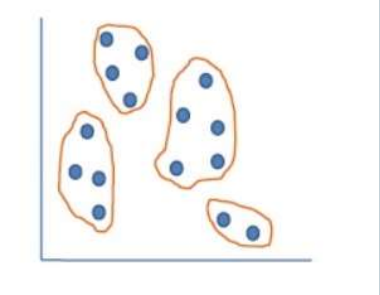
3

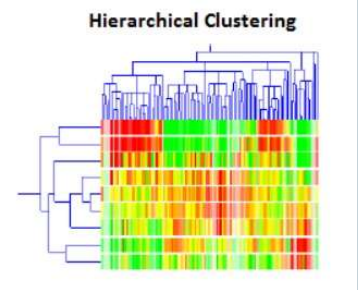
Hierarchical vs Non Hierarchical

- Non Hierarchical clustering: K-means
- El clúster no jeràrquic implica la formació de nous clústers fusionant o dividint els clústers en lloc de seguir un ordre jeràrquic.
- No segueix cap estructura d’arbre com ara l’agrupació jeràrquica. Aquesta tècnica agrupa les dades per tal de maximitzar o minimitzar alguns criteris d’avaluació. Mètode Kmeans és un exemple d’agrupació no jeràrquica.

4

Hierarchical vs non hierarchical





5

Cluster analysis: Problemàtica

- La utilització del cluster analysis presenta alguns problemes, **com són l'elecció de la mesura de similitud, la determinació del nombre de grups o clusters a retenir** (Arnold, 1979) i, en alguns casos, l'elevat nombre d'operacions a realitzar (Funkhouser, 1983).

6

Distàncies

Group 1 (Pythagorean Euclidean Distance for quantitative data and related): Euclidean Distance; Squared Euclidean Distance; Manhattan or city block metric; Absolute Value Distance; Mahalanobis Distance (Like Euclidean distance but it takes into account variable correlations and "extracts" them); Lp distance; Binary Euclidean Distance and Binary Squared Euclidean Distance (Pythagorean Euclidean Distance for binary data)

Group 2 (Angle between Profiles/Euclidean distance computed after transforming data): Chord Distance; Hellinger Distance; Species Profiles Distance; Bray-Curtis Distance (same sampling effort)

Group 3 (Statistical distances): Chi-square distance and Chi-square metric (Contingency tables, Correspondence Analysis,….).

7

Distància a escollir?

	C1	C2	C3	C4
P1	0	3	0	2
P2	2	0	1	0
P3	0	20	0	12

$$d_{EC} = \begin{pmatrix} 0 & 4.24 & 19.72 \\ 4.24 & 0 & 23.43 \\ 19.72 & 23.43 & 0 \end{pmatrix}$$

$$d_{Man} = \begin{pmatrix} 0 & 8 & 27 \\ 8 & 0 & 35 \\ 27 & 35 & 0 \end{pmatrix}$$

$$d_{BC} = \begin{pmatrix} 0 & 1 & 0.73 \\ 1 & 0 & 1 \\ 0.73 & 1 & 0 \end{pmatrix}$$

$$d_{Hell} = \begin{pmatrix} 0 & 1.41 & 0.03 \\ 1.41 & 0 & 1.41 \\ 0.03 & 1.41 & 0 \end{pmatrix}$$

8

Mesures de similaritat 1/3

Contingency table for binary data

		Object *j*		
		1	0	sum
Object *i*	1	a	b	a+b
	0	c	d	c+d
	sum	a+c	b+d	

where
a = number of double presence (occurrence)
b = number of presence-absence
c = number of absence-presence
d = number of double absence

9

Mesures de similaritat 2/3

Simple Matching Coefficient

$$s_1(x_1,x_2)=\frac{a+d}{a+b+c+d}$$

Jacard index

$$s_2(x_1,x_2)=\frac{a}{a+b+c}$$

Sorensen index

$$s_3(x_1,x_2)=\frac{2a}{2a+b+c}$$

distance: $d(x_1,x_2)=[1-s(x_1,x_2)]^{\alpha}$

10

Molt important,....

¿Quina distancia (similitud)

caracteritza les analogies i diferencies

en la situació experimental que estudiem?

11