Hierarchical Clustering

Example of centroid method

***K. Gibert*[(1)]**

[(1)]Department of Statistics and Operation Research

Knowledge Engineering and Machine Learning group
Universitat Politècnica de Catalunya, Barcelona

Hierarchical clustering (CENTROID)

0) Data matrix

	Age	Weight	cigarrettes	Hard attacks
	years	*Kg*	*Pack/week*	#
A	30	Low	High	1
B	40	High	Moderate	1
C	30	Medium	Moderate	2
F	40	Low	Low	2
J	30	High	Low	2
M	30	Medium	Low	0
P	40	High	High	0
R	30	Low	Moderate	0
S	50	High	High	2
T	40	Medium	High	2

1) Compute distances between persons (Gibert's mixed distance)

	A	B	C	F	J	M	P	R	S	T
A	0	1.2488811	1.2438452	0.8309088	1.1683081	1.2438452	0.8309088	0.63954794	1.4049911	0.90644586
B	1.2488811	0	0.8309088	1.2438452	0.90644586	1.4352059	0.63954794	0.8309088	0.8309088	1.1683081
C	1.2438452	0.8309088	0	1.3999553	1.1330574	1.0474486	1.6920323	1.0474486	1.8229634	0.7201209
F	0.8309088	1.2438452	1.3999553	0	0.7201209	1.2388093	1.5006716	1.2388093	1.2488811	1.1330574
J	1.1683081	0.90644586	1.1330574	0.7201209	0	0.97191143	1.1632721	1.5762087	1.2942033	1.2488811
M	1.2438452	1.4352059	1.0474486	1.2388093	0.97191143	0	1.2488811	1.2085946	2.2661147	1.1632721
P	0.8309088	0.63954794	1.6920323	1.5006716	1.1632721	1.2488811	0	1.2488811	0.63451207	0.97191143
R	0.63954794	0.8309088	1.0474486	1.2388093	1.5762087	1.2085946	1.2488811	0	2.2661147	1.7675694
S	1.4049911	0.8309088	1.8229634	1.2488811	1.2942033	2.2661147	0.63451207	2.2661147	0	0.7201209
T	0.90644586	1.1683081	0.7201209	1.1330574	1.2488811	1.1632721	0.97191143	1.7675694	0.7201209	0

Hierarchical clustering (CENTROID)

2) Select the more similar pair of objects (S and P)

3) Build a new class C1

4) Calculate the centroid of C1

C1: (45, High, High, 1)

5) Eliminate P and S

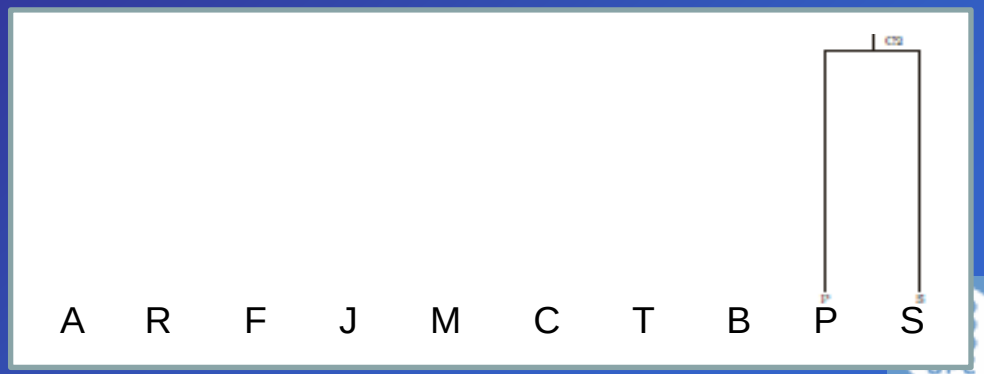
6) Start Dendrogramm

7) Compute distances between C1 and rest
(aggregation criterion)

d(ci,cj) = d(gi,gj)

8) Repeat till all elements clustered

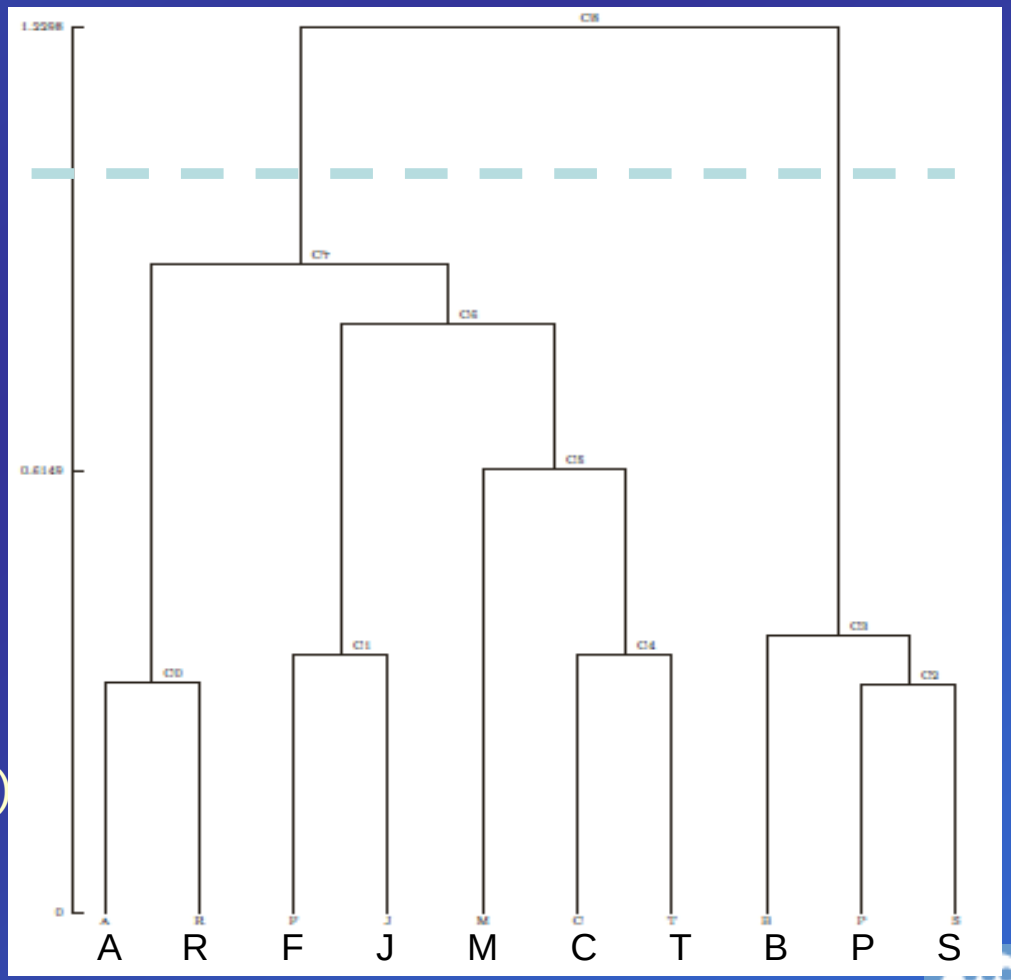
	Age	Weight	cigarrettes	Hard attacks
	years	*Kg*	*Pack/week*	*#*
A	30	Low	High	1
B	40	High	Moderate	1
C	30	Medium	Moderate	2
F	40	Low	Low	2
J	30	High	Low	2
M	30	Medium	Low	0
P	40	High	High	0
R	30	Low	Moderate	0
S	50	High	High	2
T	40	Medium	High	2



Hierarchical clustering (CENTROID)

9) Analize final dendrogram to decide number of clusters.
alternative (optimize some criteria, ex Calinski-Harabaz, Dunn, etc)

Best cut is in 2 clusters



10) Calculate 2-class partition

(intensional: P2={C7, C3}

(extensional:

C7={A,R,F,J,M,C,T}
C3={B,P,S}

11) Interpret clusters
(not evident as conceptual cluster)

12) Structural validation

Hierarchical clustering (CENTROID)

	Age	Weight	cigarrettes	Hard attacks
	years	*Kg*	*Pack/week*	#
A	30	Low	High	1
B	40	High	Moderate	1
C	30	Medium	Moderate	2
F	40	Low	Low	2
J	30	High	Low	2
M	30	Medium	Low	0
P	40	High	High	0
R	30	Low	Moderate	0
S	50	High	High	2
T	40	Medium	High	2