Gap statistic

E' stata proposta da Tibshirami et al. (2001). Cerca il numero ottimale di cluster con un confronto tra la clustering ottenuta con k gruppi e quella ottenibile da un data set ottenuto con delle repliche simulate del data set.

Misuriamo l'omogoneità del cluster C_r con l'espressione

$$h_r = \sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}\gamma_{ir}\gamma_{jr};\ \ \gamma_{ir} = \begin{cases} 1 & U_i \in C_r \\ 0 & U_i \in C \end{cases};\ \gamma_{jr} = \begin{cases} 1 & U_j \in C_r \\ 0 & U_j \in C \end{cases}$$

Il prodotto degli indicatori γ è uno se entrambe le unità appartengono a C_r

La qualità complessiva della clustering in k gruppi è misurata da

$$H_k = \sum_{r=1}^{k} w_r h_r \quad con \quad w_r = \begin{cases} \left(\frac{1}{n_r}\right) & se\ \ n_r > 0 \\ 0 & se\ \ n_r > 0 \end{cases}$$

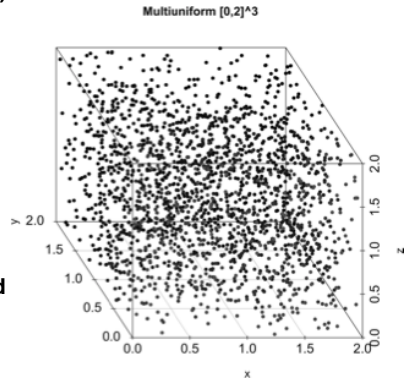
Valori piccoli di H_k indicano compattezza nei cluster.

Gap statistic/2

Si confronta la clustering ottenuta con quella che si avrebbe se la matrice dei dati fosse formata da valori casuali privi di struttura di gruppo

Per mantenere una certa coerenza con i nostri dati, i valori casuali dovrebbero avere almeno lo stesso campo di variazione e la stessa matrice di correlazione (almeno simile)

	X_1	X_2	$\cdots$	X_m
min	L_1	L_2	$\cdots$	L_m
max	U_1	U_2	$\cdots$	U_m



E' necessario ripetere la procedura un certo numero di volte. Diciamo che (ad es. 60) repliche potrebbero bastare.

Per ogni replica si calcola W_k

Gap statistic/2

Se le dissimilarità fossero delle distanze euclidee allora H_k diminuirebbe con k e sarebbe poco informativa dato che per un miglioramento basterebbe solo aumentare k.

Inoltre, la presenza di gruppi ben strutturati provocherebbe una discesa netta di H_k in corrispondenza del numero più plausibile di gruppi perché le unità si avvicinerebbero maggiormente ai loro centri.

Poiché l'euclideità della matrice delle distanze non è garantita dobbiamo trovare una alterativa

Misuriamo l'omogoneità di una clustering in k gruppi C_r con

$$W_k = \log\left(H_k\right)$$

Gap statistic/3

In base alle repliche effettuate possiamo calcoare il valore medio di Wk in caso di dati non strutturati

Il confronto della media con il valore nel data set reale si realizza con

$$G_k = \frac{\sum_{r=1}^{N} W_k^*}{N} - W_k$$

N è il numero di repliche

Il numero di cluster da valutare come migliore scelta è dato dal criterio

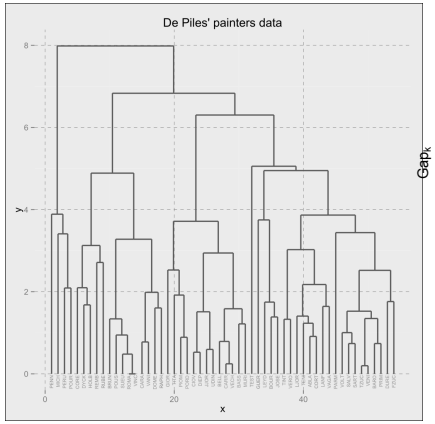
$$k\_opt = \min_{2 \le k \le k'} \left\{ G_k \ge G_{k+1} - S_{k+1}\sqrt{\frac{N+1}{N}} \right\}$$

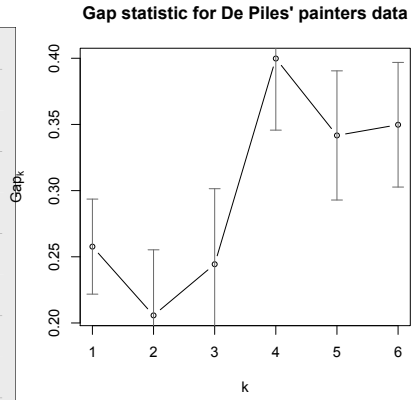
Dove S_k è la deviazione standard di W_k nelle N repliche

clusGap {cluster}

Esempio

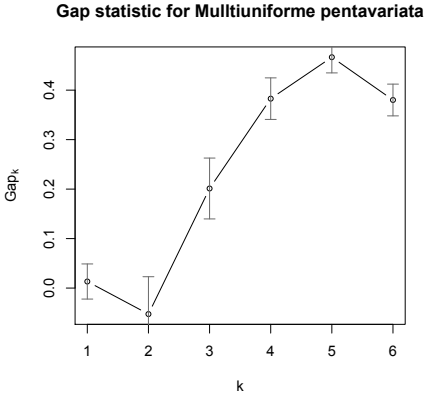
E' un data set che valuta alcuni aspetti della tenica di diversi pittori.

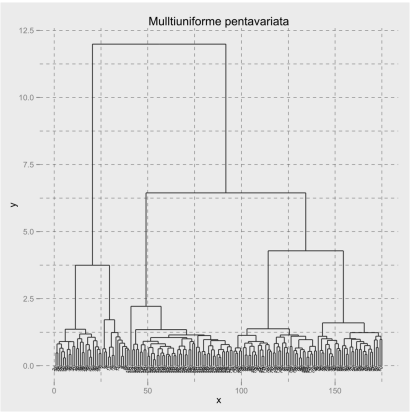




K=4 è una scelta ragionevole

Altro esempio



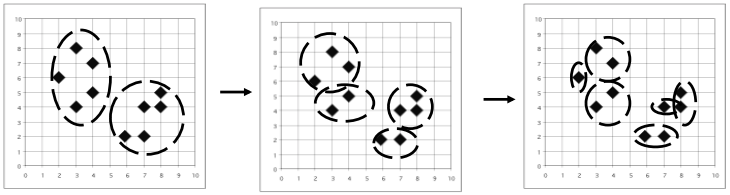


Si è individuato il corretto numero di cluster

Metodi gerarchici divisivi

La costruzione gerarchica può anche essere scissoria (o divisiva): si parte da un solo grande cluster che include tutte le entità e si arriva ad una suddivisione in n clusters di ampiezza uno.

Si procede dalla L_q alla L_{q+1} per bipartizione di uno dei cluster già in L_q.



Questo approccio ha il vantaggio di partire dal punto in cui nessun dato è disperso perché il data set è ancora intatto.

Inoltre, non è necessario procedere fino all'ultima delle suddivisioni (un cluster di due entità scomposto in due clusters di una sola entità ciascuno) dato che la qualità del risultato al livello L_q può essere controllata rispetto a quella attesa.

Metodi gerarchici divisivi/2

Nelle tecniche scissorie, il primo passaggio da $L_0=\{U_1,U_2,...,U_n\}$ ad L_1 formato da $L_1=[\{U_{i1}, U_{i2},...U_{ini}\}, \{U_{j1}, U_{j2},...U_{jnj}\}]$ consiste in una bipartizione del data set in due sottoinsiemi.

Le possibilità di scelta sono $\left(2^{n-1}-1\right)=\sum_{i=1}^{\left[\frac{n}{2}\right]}\binom{n}{i}$

Se n=5 occorre esaminare 15 suddivisioni.

$$\sum_{i=0}^{n}\binom{n}{i}=(1+1)^n=2^n;\ 2)\ \binom{n}{i}=\binom{n}{n-i} \dashrightarrow \sum_{i=1}^{\left[\frac{n}{2}\right]}\binom{n}{i}=\frac{2^n}{2}-1=2^{n-1}-1$$

Se n è grande, il numero di bipartizioni è elevato al punto che non sempre è possibile applicare tecniche scissorie.

La complessità computazionale è il maggiore ostacolo alla loro diffusione.

Metodi gerarchici divisivi/3

Si realizza una sequenza di bipartizioni (split) a partire dall'intero data set per poi fermarsi dopo un certo numero di suddivisioni

Ad ogni stadio uno dei gruppi già esistenti, diciamo C è suddiviso in due nuovi gruppi non vuoti C1 e C2.

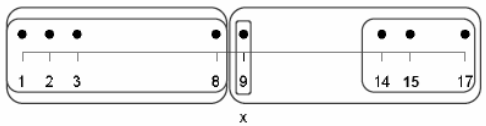
Le scelte da fare sono quindi due: quale cluster scomporre e in base a quale criterio effettuare lo split

Unità simili debbono confluire nello stesso cluster ed unità dissimili debbono essere collocate in cluster diversi

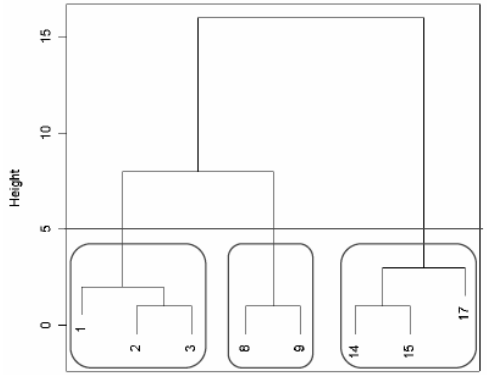
Questo principio porta a considerare la variabilità interna dei cluster da ottenere rispetto a quelli già formati.

Se si dispone solo delle distanze tra le unità si dovrà ragionare in termini di omogeneità dei cluster

Potenziali problemi



La strategia delle suddivioni dovrebbe evitare lo splitting delle unità nelc cluster centraòe



Scomposizione della varianza (univariata)

$$n\sigma^2 = \sum_{r=1}^{k}\sum_{j=1}^{n_i}\left(X_{rj}-\mu\right)^2 = \sum_{r=1}^{k}\sum_{j=1}^{n_i}\left[\left(X_{rj}-\mu_r\right)+\left(\mu_r-\mu\right)\right]^2;$$

X_{ij} valore della j-esima variabile nello i-esimo cluster

$$= \sum_{r=1}^{k}\sum_{j=1}^{n_i}\left(X_{rj}-\mu_r\right)^2 + \sum_{r=1}^{k}\sum_{j=1}^{n_i}\left(\mu_r-\mu\right)^2 + 2\sum_{r=1}^{k}\sum_{j=1}^{n_i}\left(\mu_r-\mu\right)\left(X_{rj}-\mu_r\right);$$

$$= \sum_{r=1}^{k} n_r \left[\frac{\sum_{j=1}^{n_i}\left(X_{rj}-\mu_r\right)^2}{n_r}\right] + \sum_{r=1}^{k} n_r\left(\mu_r-\mu\right)^2 + 2\sum_{r=1}^{k}\left(\mu_r-\mu\right)\sum_{j=1}^{n_i}\left(X_{rj}-\mu_r\right)$$

Zero

$$n\sigma^2 = \sum_{r=1}^{k} n_r\sigma_r^2 + \sum_{r=1}^{k} n_r\left(\mu_r-\mu\right)^2 \Rightarrow \sigma^2 = \sum_{r=1}^{k}\left(\frac{n_r}{n}\right)\sigma_r^2 + \sum_{r=1}^{k}\left(\frac{n_r}{n}\right)\left(\mu_r-\mu\right)^2$$

Si sono ottenuti due fattori additivi:

Varianza within **(media delle varianze all'interno di ciascun gruppo)**

Varianza between **(media degli scarti tra le medie di gruppo ed il centroide globale)**

Scomposizione della varianza (multivariata)

La varianza totale è data dalla somma di n prodotti esterni

$$T = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\left(X_i-\mu\right)\left(X_i-\mu\right)^t \quad dove \quad \begin{cases} X_i = \left(x_{i1},x_{i2},\ldots,x_{im}\right) \; riga\ della\ matrice\ dei\ dati \\ \mu = \left(\mu_1,\mu_2,\ldots,\mu_m\right) \; centroide\ della\ matrice\ dei\ dati \end{cases}$$

Se esistono k gruppi di numerosità $n_1,n_2,\ldots,n_k$ ognuno con il proprio centroide $\mu_1, \mu_2, \ldots, \mu_k$, la devianza totale è espressa da

$$nT = \sum_{r=1}^{k}\sum_{i=1}^{n}\gamma_{i,r}\left(X_i-\mu\right)\left(X_i-\mu\right)^t =$$

$$= \sum_{r=1}^{k}\sum_{i=1}^{n}\gamma_{i,r}\left[\left(X_i-\mu_r\right)-\left(\mu-\mu_r\right)\right]\left[\left(X_i-\mu_r\right)-\left(\mu-\mu_r\right)\right]^t, \quad \mu_r = \left(\mu_{1r},\mu_{2r},\ldots,\mu_{mr}\right)$$

Dove γ_{ir} è la funzione indicatore che è uno se l'unità i appartiene al cluster r e zero altrimenti

Scomposizione della varianza/2

Lo sviluppo dei prodotti comporta

$$nT = \sum_{r=1}^{k}\sum_{i=1}^{n}\gamma_{i,r}(X_i-\mu_r)(X_i-\mu_r)^t - \sum_{r=1}^{k}\sum_{i=1}^{n}\gamma_{i,r}(X_i-\mu_r)(\mu-\mu_r)^t - \\ - \sum_{r=1}^{k}\sum_{i=1}^{n}\gamma_{i,r}(\mu-\mu_r)(X_i-\mu_r)^t + \sum_{r=1}^{k}\sum_{i=1}^{n}\gamma_{i,r}(\mu-\mu_r)(\mu-\mu_r)^t$$

Isoliamo le quantità legate ad un solo indice

$$nT = \sum_{r=1}^{k}\sum_{i=1}^{n}\gamma_{i,r}(X_i-\mu_r)(X_i-\mu_r)^t - \sum_{r=1}^{k}\left[\sum_{i=1}^{n}\gamma_{i,r}(X_i-\mu_r)\right](\mu-\mu_r)^t \\ - \sum_{r=1}^{k}(\mu-\mu_r)\sum_{i=1}^{n}\left[\gamma_{i,r}(X_i-\mu_r)^t\right] + \sum_{r=1}^{k}\sum_{i=1}^{n}\gamma_{i,r}(\mu-\mu_r)(\mu-\mu_r)^t \\ = \sum_{r=1}^{k}\sum_{i=1}^{n}\gamma_{i,r}(X_i-\mu_r)(X_i-\mu_r)^t + \sum_{r=1}^{k}\sum_{i=1}^{n}\gamma_{i,r}(\mu-\mu_r)(\mu-\mu_r)^t$$

Le espressioni in parentesi quadre sono nulle in quanto somme degli scarti dalle medie aritmetiche all'interno dei gruppi.

Scomposizione della varianza/3

In estrema sintesi si è ottenuta la scomposizione della varianza

$$T = \left(\frac{1}{n}\right)\sum_{r=1}^{k}\sum_{i=1}^{n}\gamma_{i,r}(X_i-\mu_r)(X_i-\mu_r)^t + \left(\frac{1}{n}\right)\sum_{r=1}^{k}\sum_{i=1}^{n}\gamma_{i,r}(\mu-\mu_r)(\mu-\mu_r)^t$$

$$Totale = Within + Between \Rightarrow T = W + B$$

La suddivisione del data set in gruppi non ha effetto se le medie per gruppo sono tutte uguali tra di loro (e dunque uguali alla media totale)

Se le medie parziali sono diverse allora i gruppi sono una fonte di variabilità che è tanto più forte quanto più cresce il divario rispetto al centroide dei dati.

Ogni suddivisione del data set in gruppi varierà entrambi i termini della scomposizione

Dato che la varianza totale è fissa: se W aumenta allora B deve diminuire e viceversa

Metodo Edwards e Cavalli-Sforza

Questi due studiosi hanno proposto di organizzare le scissioni dei gruppi in modo da rendere minima, per quella suddivisione, la traccia della matrice di devianze-codevianze "within" rispetto ai gruppi

$$T = \left(\frac{1}{n}\right)\sum_{r=1}^{k}\sum_{i=1}^{n}\gamma_{i,r}(X_i-\mu_r)(X_i-\mu_r)^t$$

La stessa strategia si avrebbe minimizzando la traccia di B sebbene la minimizzazione di Tr(W) sia più semplice

Il metodo più sicuro per trovare la suddivisione ottima è di esaminarle tutte e scegliere quella a cui è associato il minor valore di Tr(W).

L'esame a tappeto non è però praticabile: se un computer fosse in grado di esaminare 1 milione di suddivisioni al secondo per effettuare il primo passaggio con n=50 occorrerebbero circa 34 anni.

Nessun computer è oggi in grado di realizzare l'esame di un milione di suddivisioni in un secondo

Semplificazione

Il fatto è che, non tutte le suddivisioni sono interessanti e gran parte del tempo sarebbe sprecato ad esaminare clustering improponibili.

Un primo elemento di semplificazione deriva dal legame tra il quadrato della distanza euclidea tra coppie di entità e la traccia della matrice within.

Ricordando che, nonostante le diverse dimensioni, $Tr(XX^t)=Tr(X^tX)$ si ha

$$Tr(W) = \sum_{r=1}^{k}\sum_{i=1}^{n}\gamma_{i,r}(X_i-\mu_r)(X_i-\mu_r)^t = \sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{i,r}\gamma_{j,r}(X_i-\mu_r)^t(X_j-\mu_r) \\ = \sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{i,r}\gamma_{j,r}d_{ij}^2$$

che consente di valutare gli split delle unità e nel contempo mantenersi vicini alla semplice clustering gerarchica.

NB basarsi sulle distanze euclidee implica che le tecniche scissorie possono applicarsi anche con altre misure di distanza.

Esempio

	X_1	X_2
A	1	2
B	4	6
C	3	7
D	2	5
E	5	4
F	3	9

	B	C	D	E	F
A	25	29	10	20	53
B		2	5	5	10
C			5	13	4
D				10	17
E					29

x_1	$(x_1-\bar{x}_1)$	$(x_1-\bar{x}_1)^2$	x_2	$(x_2-\bar{x}_2)$	$(x_2-\bar{x}_2)^2$
1	−1.5	2.25	2	−3	9
4	1.5	2.25	6	1	1
3	0.5	0.25	7	2	4
2	−0.5	0.25	5	0	0
T10	0.0	5.00	20	0	14

$\bar{x}_1 = 2.5; \bar{x}_2 = 5$

Tr(W)=5+14=19

La traccia della matrice per lo split: (A,B,C,D) e (E,F) può essere ricavata dalla matrice delle distanze euclidee al quadrato:

var(w1)=25+29+10+2+5+5/4=76/4=19; var(w2)=29/2=14.5

Da notare che Tr(w)=19+14.5=33.5, Tr(T)=237/6=39.5 e che Tr(B)=39.5-33.5=6.

x_1	$(x_1-\bar{x}_1)$	$(x_1-\bar{x}_1)^2$	x_2	$(x_2-\bar{x}_2)$	$(x_2-\bar{x}_2)^2$
5	1	1	4	−2.53	6.25
3	−1	1	9	2.5	6.25
8	0	2	13	0	12.5

var(w2)=2+12.5=14.5

Esempio (continua)

La matrice delle distanze euclidee al quadrato ha tutte le informazioni utili a valutare ogni possibile suddivisione.

Per n=6 sono 31=2^5-1.

Lo splitting più efficace risulta essere {A,D,E}, {B,C,F} che induce il minimo incremento nella dispersione "within" cioè da zero a 18.67

Per determinare tale suddivisione ottimale è stato necessario esaminare tutte le 31 possibilità.

A questo punto la devianza totale: 39.5 è divisa in

Tr(w)=18.67; Tr(B)=20.83

split	C_1	C_2	$Tr(w_1)$	$Tr(w_2)$	Tr(w)
1	A	{B,C,D,E,F}	0	20	20
2	B		0	32.2	32.2
3	C		0	36.8	36.8
4	D		0	43.4	43.4
5	E		0	32	32
6	F		0	22.8	22.8
7	A,B		12.5	19.5	32
8	A,C		14.5	19	33.5
9	A,D		5	15.25	20.25
10	A,E		10	10.75	20.27
11	A,F		26,5	10	36.5
12	B,C	*Non in C1*	1	34.75	35.75
13	B,D		2.5	37	39.5
14	B,E		2,5	32.5	35
15	B,F		5	21.75	26.75
16	C,D		2.5	35.5	38
17	C,E		6.5	30	36.5
18	C,F		2	18.75	20.75
19	D,E		5	30.75	35.75
20	D,F		8.5	23.5	32
21	E,F		14.5	19	33.5
22	A,B,C		18.67	18.67	37.33
23	A,B,D		13.33	15.33	28.67
24	A,B,E		16.67	8.67	25.33
25	A,B,F		29.33	9.33	38.67
26	A,C,D		14.67	14.67	29.33
27	A,C,E		20.67	10.67	31.33
28	A,C,F		28.67	6.67	35.33
29	{A,D,E}{B,C,F}		13.33	5.33	18.67
30	A,D,F		26.67	6.67	33.33
31	A,E,F		34	4	38

Esempio (continua/2)

Si può tentare una ulteriore suddivisione in uno dei due clusters con due sub matrici di quella originale.

Cluster 1			Cluster 2		
	D	E		C	F
A	10	20	B	2	10
D		10	C		4

Per rendere i calcoli più praticabili, MacQueen (1967) ha suggerito di considerare, ad ogni singolo livello, solo il cluster con maggiore devianza.

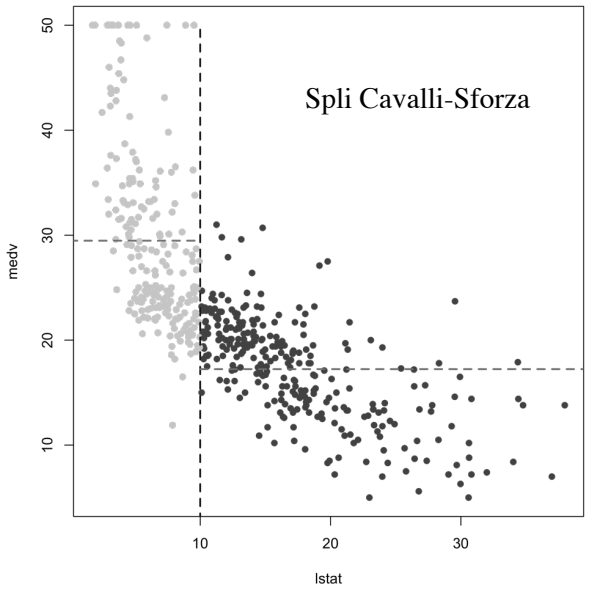
	$Tr(w_1)$	$Tr(w_2)$	$Tr(w)$		$Tr(w_1)$	$Tr(w_2)$	$Tr(w)$
$A\{D,E\}$	0	5	5	$B\{C,F\}$	0	2	2
$D\{A,E\}$	0	10	10	$C\{B,F\}$	0	5	5
$E\{A,D\}$	0	5	5	$F\{B,C\}$	0	2	1

Tra tutti i possibili split in tutti i clusters quello più efficiente è nel cluster 2 con {F}, {B,C} che porta a {A,E,D}, {B,C}, {F} con Tr(w1)=13.33, Tr(w2)=1 e Tr(B)=25.17.

Lo splitting di C1 può avvenire in due modi {A}, {D,E} e {E}, {A,D} bisognerà quindi seguire due possibilità.

Il metodo Cavalli-Sforza, anche se semplificato, è molto dispendioso e si può realizzare solo con poche unità.

Applicazione: Boston house data set



Metodo divisivo di Macnaughton-Smith (Diana)

Al fine di aggirare l'immenso ostacolo delle operazioni di splitting occorre procedere per via approssimata e con iterazioni.

L'idea è di attuare la bipartizione di un cluster collocando una unità alla volta nel gruppo nuovo scegliendo tra quelle più marginali del gruppo da dividere.

La prima unità trasferita nel gruppo nuovo è quella la cui distanza media è maggiore e che quindi contribuiva di più alla dispersione del gruppo originale

	B	*C*	*D*	*E*	*F*
A	25	29	10	20	53
B		2	5	5	10
C			5	13	4
D				10	17
E					29

$\bar{d}(A) = (25+29+10+20+53)/5 = 25.4$
$\bar{d}(B) = (25+2+5+5+10)/5 = 9.4$
$\bar{d}(C) = (29+2+5+13+4)/5 = 10.6$
$\bar{d}(D) = (10+5+5+10+17)/5 = 7.4$
$\bar{d}(E) = (20+5+13+10+29)/5 = 15.4$
$\bar{d}(F) = (53+10+4+17+29)/5 = 22.6$

L'entità più remota dal resto del cluster è la "A" che ha una dissimilarità media maggiore di ogni altra unità. Quindi, nel secondo gruppo deve essere collocata la "A" per cui la suddivisione è {A} e {B,C,D,E,F}.

Diana/2

Resta da vedere ora se l'unità trasferita non trascini con se qualche altra entità che potrebbe risultare più prossima al nuovo gruppo (che al momento contiene solo la A) piuttosto che al vecchio gruppo.

Dobbiamo quindi confrontare, per ciascuna delle unità che rimangono nel vecchio gruppo, la distanza o dissimilarità media dal vecchio e dal nuovo gruppo

	B	*C*	*D*
$\bar{d}_O$	$(2+5+5+10)/4 = 5.5$	$(2+5+13+4)/4 = 6$	$(5+5+13+4)/4 = 6.75$
$\bar{d}_N$	25	29	10
	−19.5	−23	−3.25

	E	*F*
$\bar{d}_O$	$(5+13+10+29)/4 = 13$	$(10+4+17+29)/4 = 15$
$\bar{d}_N$	20	53
	−7	−38

O=old, N=new

Nessuna altra entità segue la A che quindi rimane da sola a questo livello.

N.B. Il primo splitting è dunque meno efficiente del passo Cavalli Sforza.

Diana/3

A questo punto ci concentriamo sul cluster {B,C,D,E,F} per verificare se c'è un'altra entità marginale (l'altro, a questo punto, contiene solo la A).

L'entità candidata è la F, con dissimilarità media di 15, superiore alle altre. Vediamo se qualche altra unità la accompagna

	B	*C*	*D*	*E*
$\bar{d}_O$	$(2+5+5)/3 = 4$	$(2+5+13)/3 = 6.61$	$(5+5+10)/3 = 6.67$	$(5+13+10)/3 = 9.33$
$\bar{d}_N$	10	4	17	29
	−6	2.67	−10.33	−19.67

L'unità C deve cambiare cluster in quanto è più vicina al nuovo gruppo che al vecchio. Effettuiamo lo spostamento e ricontrolliamo le dissimilarità medie

	B	*D*	*E*
$\bar{d}_O$	$(5+5)/2 = 5$	$(5+10)/2 = 7.5$	$(5+10)/2 = 7.5$
$\bar{d}_N$	$(10+2)/2 = 6$	$(5+17)/2 = 11$	$(13+29)/2 = 21$
	−1	−3.5	−13.5

Poiché le differenze sono tutte negative è da ritenere che non vi siano altri spostamenti utili a questo livello.

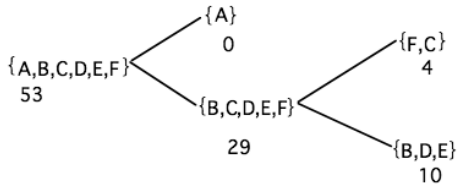
Diana/4

Una ulteriore suddivisione può essere tentata operando sul cluster che ora risulta più numeroso: {B,D,E} (in verità occorre esaminali tutti).

Un'occhiata alla tabella delle distanze nei cluster attuali rivela che sia la D che la E hanno una distanza media che è maggiore o uguale delle altre

Con tre sole unità non è il caso di proporre uno splitting con distanze medie così simili.

L'albero delle divisioni è quindi



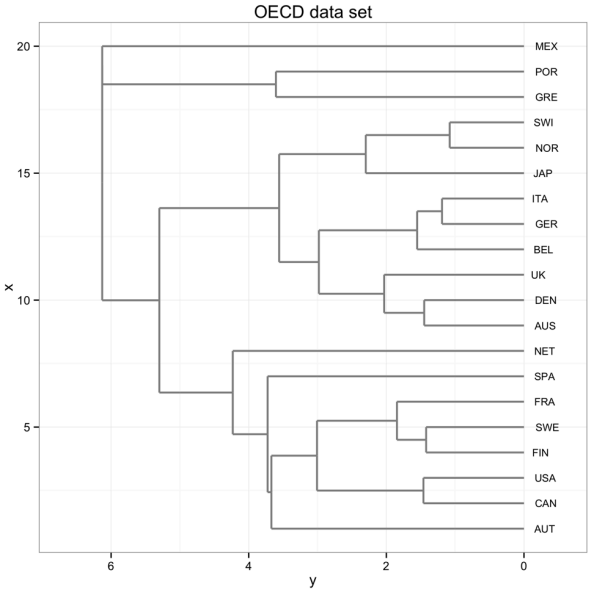
Il numero sotto il cluster è il diametro del cluster, cioè la distanza massima interna al cluster.

Esempio

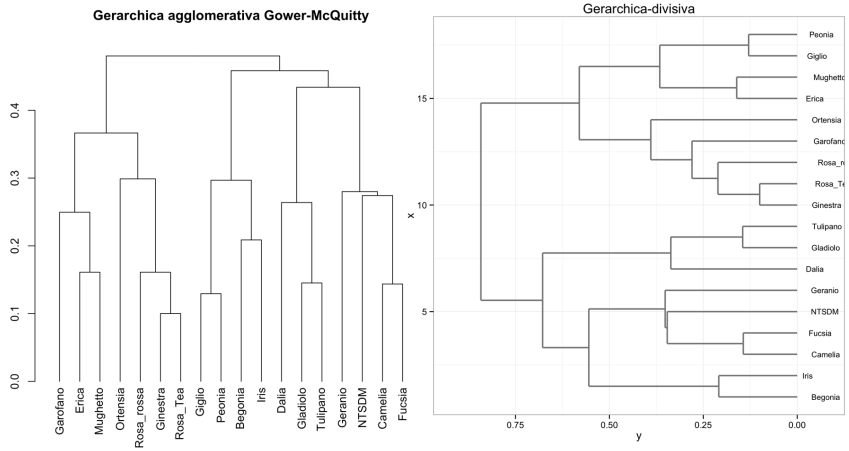
Paesi dell'OECD per alcune variabili macroeconomiche.

Si può notare la presenza di tre gruppi tra cui il cluster marginale formato da Portogallo Grecia

Il Mexico è un outlier



Flower data set (Everitt)



Ci sono analogie e ci sono differenze. Si tratta di due angolature diverse ed entrambe possono contribuire a chiarire un problema di classificazione. La aggregativa va meglio se ci aspettano molti gruppi; la divisiva è da preferire se si cercano pochi cluster.

Mon.A

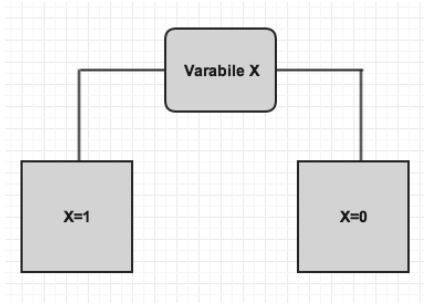
Kaufman e Rousseeuw dedicano uno spazio particolare alle matrici di dati in cui siano presenti solo variabili binarie.

In questo caso NON si procede alla costruzione della matrice delle distanze.

Ad ogni stadio si sceglie un cluster, una delle variabili binarie e si procede alla bipartizione.

Il processo continua fino a quando non si raggiunge una regola di stop oppure si pervien ad una clustering formata da tutti singoletti.

La strategia è gerachicha perché i livell di bipartizione si annidano ed è monotetica perché si usa una variabile alla volta per riallocare le unità del gruppo prescelto.



Mon.A/2

La scelta cruciale riguarda la variabile secondo la quale dividere un cluster.

Dovrebbe essere quella più vicina a tutte le altre variabil secondo una certa misura di associazione.

Variabile 2

Variabile 1

	1	0	
1	a	b	$a+b$
0	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d$

Michael	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$
Yule _1	$\frac{ad-bc}{ad+bc}$
Yule _2	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$

I coefficienti a sinitra variano tra -1 ed 1 e rimangono definiti in presenza di una colonna o una riga di zeri

Esempio

Soggetti	X1	X2	X3	X4	X5	X6	X7	X8
Biondi	0	1	1	0	0	0	0	0
Aranci	0	0	0	0	1	0	0	0
Lavandi	0	1	1	0	1	0	0	1
Verdi	0	0	0	1	0	1	0	0
Gialli	0	0	1	1	1	0	0	0
Violetti	0	1	1	1	1	0	0	1
Neri	0	0	0	1	1	1	0	0
Bianchi	1	0	0	0	0	1	1	0
Rossi	1	0	0	1	1	1	1	1
Bruni	1	1	1	0	0	1	1	0
Indaci	1	1	0	0	1	1	0	0
Rosati	1	1	0	1	1	1	0	1
Mori	1	1	0	1	0	1	1	0
Celesti	1	1	1	0	1	0	0	0
Marroni	1	1	1	0	0	0	1	0
Rosei	1	1	1	1	1	0	0	1
Azzurri	1	1	1	0	1	1	0	0
Cremisi	1	0	0	1	0	1	1	1

Associazione Yule_2

	X1	X2	X3	X4	X5	X6	X7	X8
X1	0.00	0.56	-0.23	-0.23	-0.35	0.74	1.00	0.18
X2	0.56	0.00	0.88	-0.63	0.14	-0.50	-0.33	0.18
X3	-0.23	0.88	0.00	-0.60	0.23	-0.93	-0.47	0.00
X4	-0.23	-0.63	-0.60	0.00	0.23	0.43	0.00	0.82
X5	-0.35	0.14	0.23	0.23	0.00	-0.50	-0.92	0.67
X6	0.74	-0.50	-0.93	0.43	-0.50	0.00	0.75	-0.17
X7	1.00	-0.33	-0.47	0.00	-0.92	0.75	0.00	0.00
X8	0.18	0.18	0.00	0.82	0.67	-0.17	0.00	0.00

Cerchiamo la variabile più centrale

La somma di valori con segno innesca effetti compensativi non sempre ragionevoli. Calcoliamo le somme solo su valori positivi

X1	X2	X3	X4	X5	X6	X7	X8
2.477	1.755	1.113	1.478	1.263	1.918	1.750	1.838

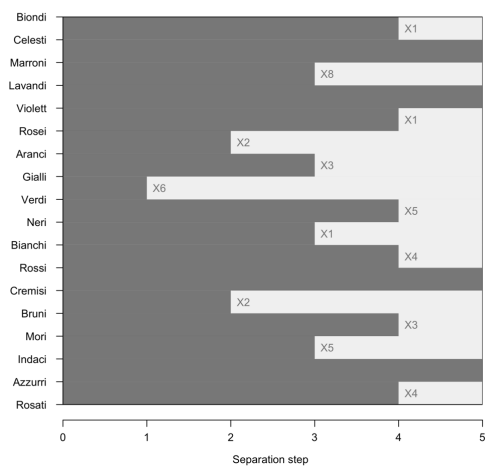
La variabile candidata per lo splitting è la X1

Il prossimo split sarà per la X4 (3ª) nel secondo cluster (x1=0)

	Ac1	Ac2
[1,]	1.538	2.0
[2,]	1.200	2.2
[3,]	1.733	1.2
[4,]	1.538	1.4
[5,]	0.963	1.0
[6,]	0.738	0.0
[7,]	1.691	3.0

Mon.A (package cluster)

Cognomi a colori



Misura di associazione del comando mona

$$\phi_{ij} = |ad - bc|$$

Il separation step è il valore del coefficiente prima di decidere lo split.

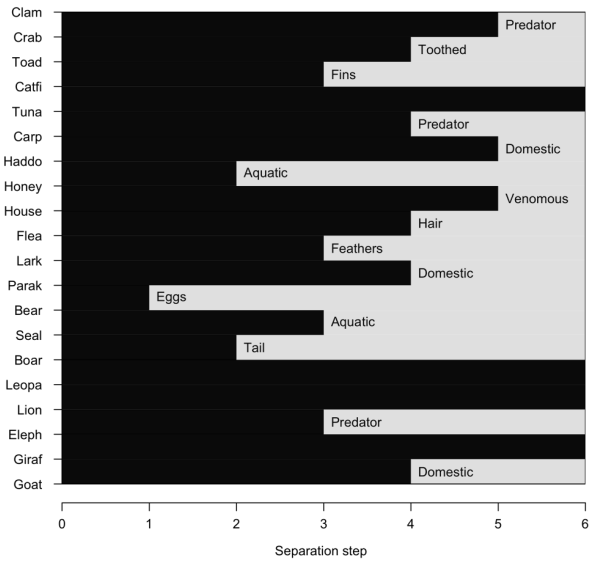
Le unità sono nell'ordine trovato dall'algoitmo.

La variabile responsabile dello split è riportata in capo alla barra.

La lunghezza della barra è data dal numero di split effettuati per arrivarci

Altro esempio

Binary variables



Se la barra arriva fino al margine allora il cluster corrispondente non può essere frazionato.

Classificazione ammissibile

La classificazione perfetta è troppo rigida dato che esclude casi invece del tutto accettabili

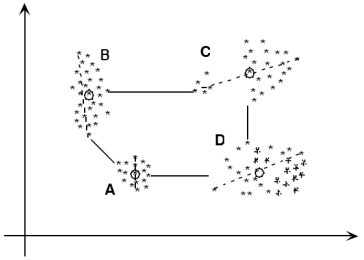


Figure 4: Admissible, but non perfect clustering

E' opportuno disporre di una definzione meno vincolante.

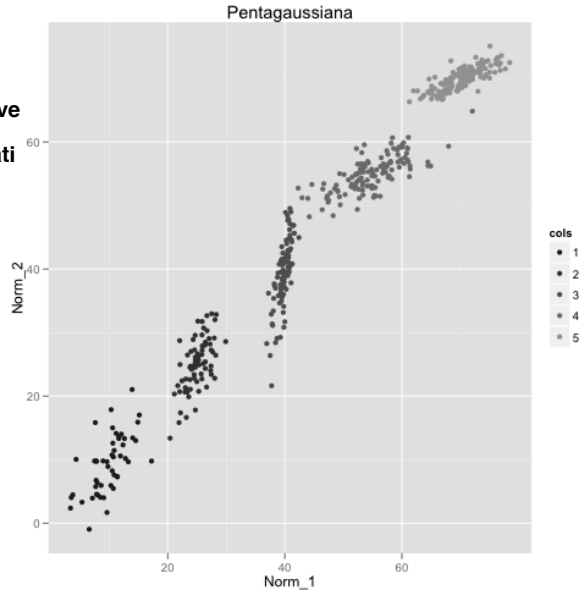
Una classificazione o clustering è ammissibile se in ogni cluster si trova una unità tale che le unità esterne al cluster distano da essa più delle unità interne al cluster

$$\min_{\gamma_i = r}\left\{d(U_i,\mu_r);\ i = 1,2,\cdots,n\right\} \geq \max_{\gamma_i \neq r}\left\{d(U_i,\mu_r);\ i = 1,2,\cdots,n\right\}$$

Il problema della cluster analysis non è più di cercare una gerarchia di gruppi, ma di individuare le unità leader dei vari gruppi

Esempio

In situazioni come quella qui presentata ogni algoritmo ammissibile deve proporre in prima istanza una classificazione dei dati in 5 gruppi



Raggruppamento iterativo

Le procedure che rientrano in questa seconda classe hanno un concetto "metrico" del cluster, lo vedono cioè come un insieme di punti che sono più o meno vicini in uno spazio metrico vicini.

All'interno del cluster si individua il cosiddetto "centroide" che costituisce il punto di massimo addensamento delle unità ovvero il punto verso cui le unità confluirebbero se non ci fossero forze differenzianti (la variabilità o la eterogeneità) a tenerli distinti.

In genere il centroide è dato dal vettore delle medie calcolate su tutte e solo le unità incluse nel cluster, ma sono possibili altre definizioni (ad esempio il vettore delle mediane).

Fissato il numero di cluster e scelta una configurazione iniziale dei centroidi si procede alla escussione delle unità assegnandole, in base ad un qualche criterio, ai cluster cui sono più vicine.

A questo punto si ricalcolano i centroidi e si effettua una nuova escussione.

Il processo termina quando nessuna unità cambia di cluster.

Scelte necessarie

La strategia di raggruppamento iterativo si precisa con delle scelte e si articola in varie fasi. Le scelte riguardano

Il numero di cluster che, a differenza dei metodi gerarchici, qui deve essere fissato di volta in volta

Il criterio in base al quale una partizione può essere giudicata preferibile rispetto ad un'altra

Il tipo di "passo" cioè lo schema di riallocazione delle unità tra un cluster e l'altro. Infatti qui le appartenenze delle unità possono essere modificate.

Le fasi cruciali del raggruppamento iterativo però sono sempre due:

La prima consiste nella definizione dei centroidi iniziali che costituiranno i poli del primo addensamento delle unità. Tale scelta può determinare il successo o l'insuccesso del metodo.

La seconda fase è costituita dal processo di assegnazione e rilocazione delle unità ed è caratterizzata dalle scelte del criterio e del passo

Il numero di partizioni

I metodi di raggruppamento iterativi presuppongono che il numero di cluster k sia già fissato, o meglio, lo considerano un parametro della procedura, che può quindi cambiare da prova a prova, ma che rimane fisso nella data applicazione.

Il modo più efficiente di suddividere n unità in k cluster è quella di cercare l'ottimo del criterio prescelto valutandolo su tutte le possibili partizioni di n oggetti in k gruppi

Come si è già visto, tale possibilità è solo teorica, in quanto realizzabile solo per valori molto piccoli di n. Ad esempio esempio per n=20 e k=3 si devono esaminare 580'606'446 partizioni

Ne consegue che la ricerca della partizione ottimale non può che avvenire esaminando un numero limitato di partizioni possibili.

E' questo uno dei limiti del raggruppamento iterativo: tranne che per le situazioni in cui sono coinvolte pochissime unità non si potrà essere sicuri che la soluzione trovata sia veramente quella ottima.

Raggrupamento iterativo: schemi generali

L'obiettivo comune è di determinare la clustering ammissibile che sia più efficace rispetto alla omogeneità interna dei cluster ed al loro isolamento esterno.

Esistono diverse strategie per questa finalità, ma ne studieremo solo due

Partitining around medoids (PAM)

Si adoperano le distanze tra le unità. I centroidi coincidono con k unità tra le n considerate

Metodo delle k-medie

Si usa la matrice dei dati limitatamente alle variabili su scala a rapporti ed intervalli. Talvolta anche le variabili binarie. I centroidi sono ottenuti come vettori delle medie dei valori osservati nei k cluster

P.A.M (algoritmo)

0) Si scelgono k unità (anche a caso) distinte e distanti che possano agire da centroidi dei cluster

1) Si forma la clustering assegnando le varie unità al cluster il cui centroide è più vicino

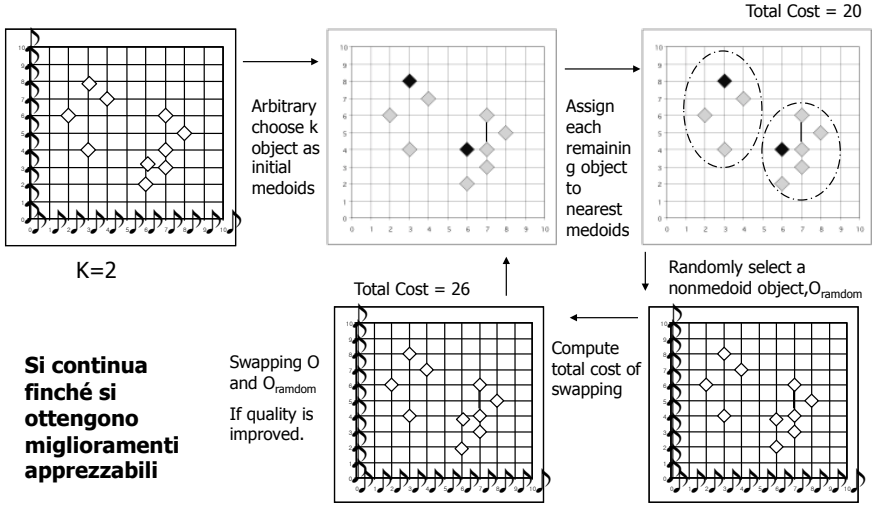
2) Si ridefiniscono i centroidi in base alla nuova articolazione dei gruppi

$$M_r = \left\{ U_r \Big| \sum_{\gamma_i = r} d(U_r, U_i) = \text{minimo} \right\}$$

3) Si valuta la qualità Q_k della clustering ottenuta. Se è valida si fermano le elaborazioni altrimenti riprendono dal punto 1.

$$Q_k = \sum_{r=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{ir} \gamma_{jr} d_{ij}$$

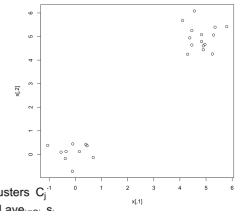
Funzionamento della tecnica PAM

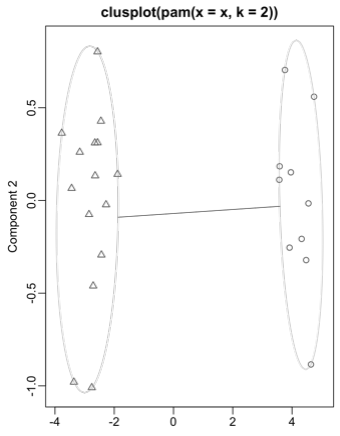


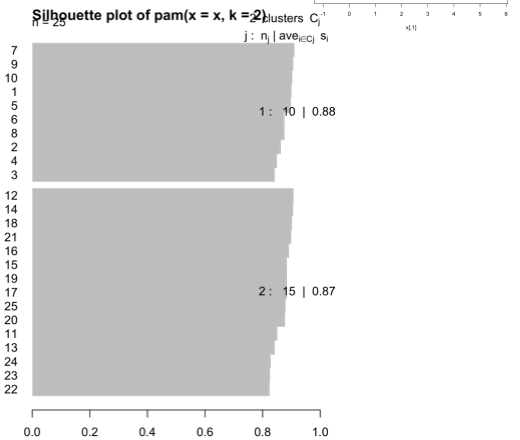
Si continua finché si ottengono miglioramenti apprezzabili

Esempio

La soluzione in questo caso è soddisfacente tanto che non cambia se si modificano i centroidi iniziali

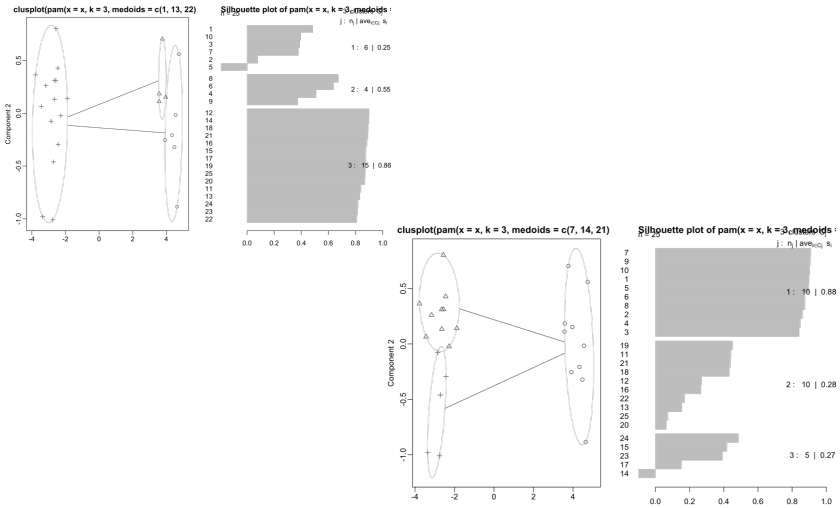






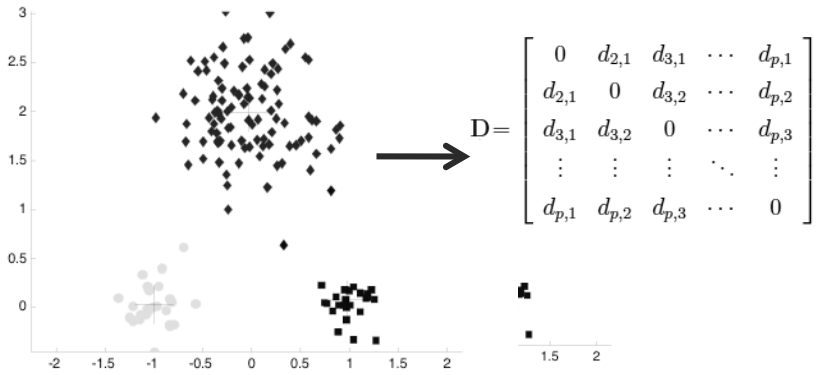
Esempio (continua)

Se si richiedono 3 cluster la soluzione si modifica secondo i centroidi iniziali



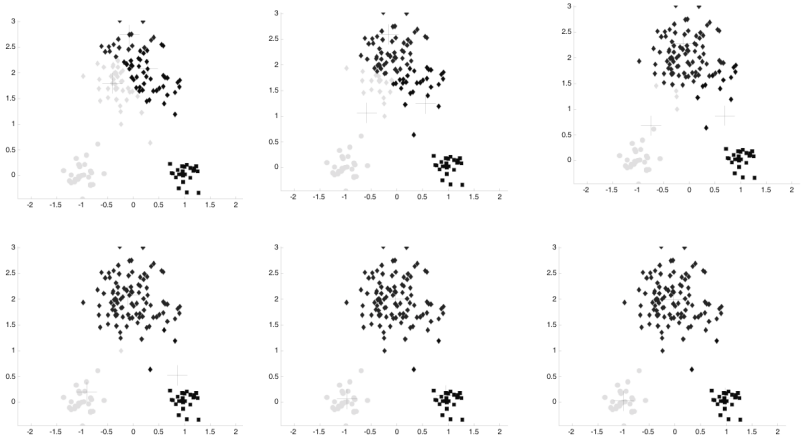
Importanza dei centroidi iniziali

Se si richiedono 3 cluster la soluzione si modifica secondo i centroidi iniziali



Importanza dei centroidi iniziali/2

La scelta a caso di k unità non da molte garanzie. Anche la scelta dei centroidi come unità occupanti posizioni particolari nel data set



Le scelte più efficaci sono quelle derivate da una applicazione di una tecnica meno impegnativa: gerarchica (agglomerativa o divisiva) oppure applicare in via preliminare la PAM ad un subset di addestramento.

P.A.M (medoidi iniziali)

Fra i tanti metodi disponibili discutiamo una strategia che ha come finalità la copertura uniforme dello spazio delle unità descritto dalla matrice delle distanze.

Il primo medoide coincide con l'unità meno periferica rispetto a tutte le altre ovvero l'unità per la quale è minore il totale delle distanze

$$M_1 = \left\{ U_r \middle| \sum_{i=1}^{n} d(U_r, U_i) = \text{minimo} \right\}$$

Indichiamo con M= {M_1, M_2, …, M_s} l'insieme delle unità già scelte come medoidi e sia s < k.

L'unità che si va ad aggiungere agli altri medoidi deve essere distanziata da essi al fine di descrivere quelle parti del data set non ancora ben rappresentate.

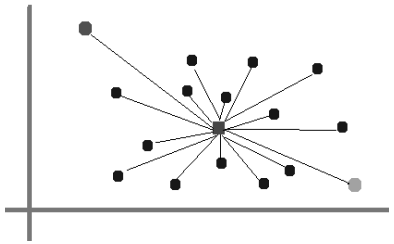
P.A.M (medoidi iniziali)/2

Per riassumere la distanza che separa l'unità entrante dai medoidi già in M si può optare per la minima distanza massima.

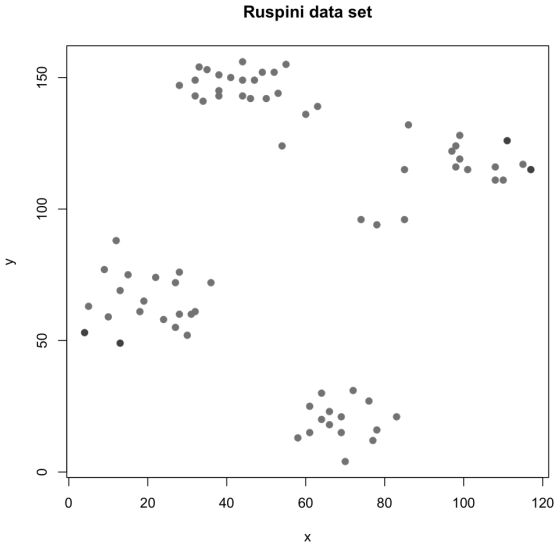
Si potrebbe anche scegliere il nuovo medoide come l'unità che ha la massima distanza media o mediana da quelle ora in M. Tale opzione dovrebbe frenare la tendenza a costruire cluster nei pressi di valori remoti.

Il metodo Kennard-Stone trova i primi due medoidi come le unità poste a più grande distanza reciproca. Quelli successivi si possono ottenere con lo stesso principio oppure uno dei precedenti

E' evidente che se si hanno informazioni su qualcuno dei medioidi potranno essere inserite e limitare la ricerca degli altri medoidi solo per quelli mancanti.

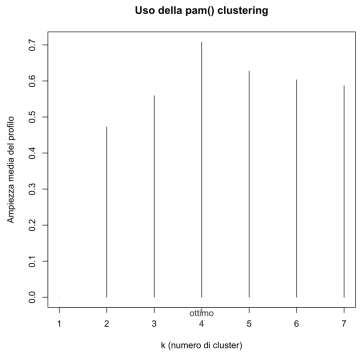


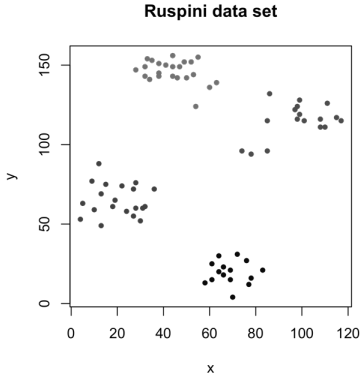
Esempio



Pur avendo provato quattro metodi diversi per determinare i medoidi le unità prescelte sono sempre quelle in rosso

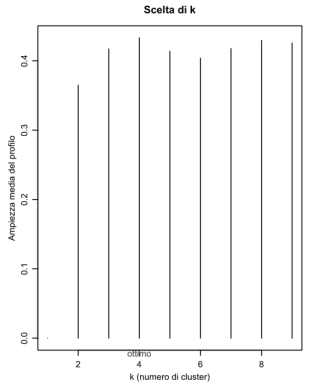
Esempio (continua)

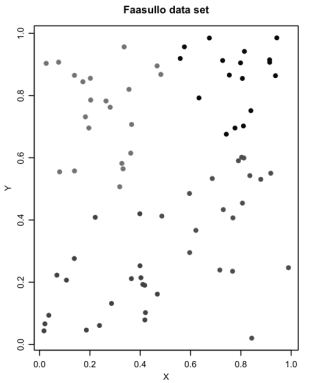




Se il data set è ben strutturato, la soluzione data dalla tecnica PAM è sempre soddisfacente

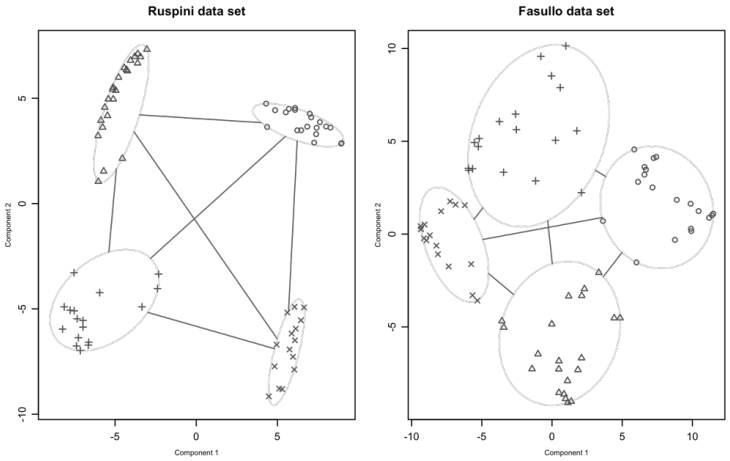
Esempio di gruppi non-gruppi





Se il data set è ben strutturato, la soluzione data dalla tecnica PAM è sempre soddisfacente

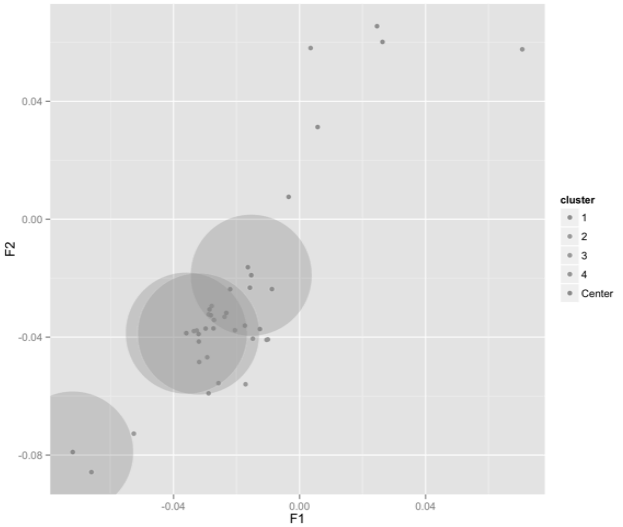
Confronto con il CusPlot



Con due sole variabili si può sfruttare la rappresentazione su piano cartesiano (usando rgl si può lavorare in 3d)

I gruppi del Ruspini sono molto più coesi ed isolati rispetto a quelli del fasullo.E' questo che convince della loro autenticità.

Applicazione: Coffee data set

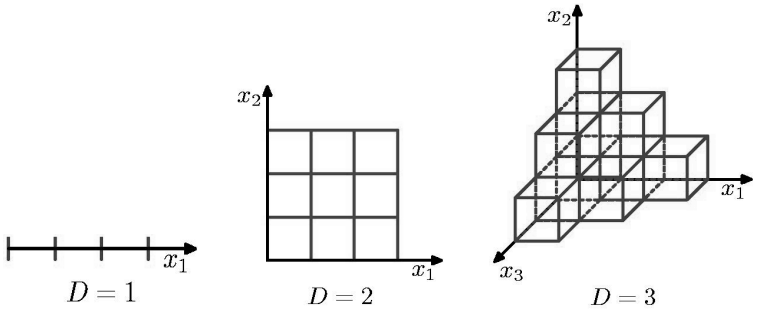


Il grafico è stato ottenuto usando la clustering su 4 gruppi e le prime due coordinate principali.

Meriti della tecnica PAM

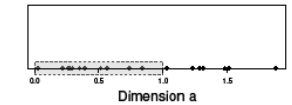
- **Si basa sulla matrice delle distanze e può quindi essere utilizzato in una vasta gamma di contesti, anche dove il concetto di metricità non è appropriato.**
- **I centroidi dei cluster (unità leader) sono delle entità proprie del data set e non unità artificiali sintetizzate dall'algoritmo di classificazione. In aggiunta si collocano al centro dei rispettivi cluster caratterizandone la natura**
- **L'aggiornamento dei cluster non richiede di ricalcolare le distanze perché si usano quelle originali.**
- **E' poco sensibile ai valori remoti fintanto che la funzione per calcolare le distanze le mantiene in un intervallo definito.**
- **L'algoritmo PAM non impone una struttura ai dati come fanno le tecniche gerarchiche**

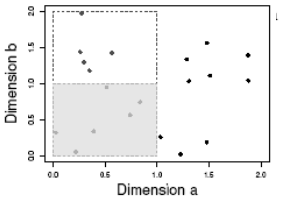
Curse of Dimensionality



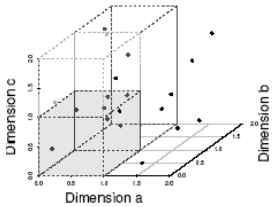
The Curse of Dimensionality

- **Data in only one dimension is relatively packed**
- **Adding a dimension "stretch" the points across that dimension, making them further apart**
- **Adding more dimensions will make the points further apart—high dimensional data is extremely sparse**
- **Distance measure becomes meaningless—due to equi-distance**





(b) 6 Objects in One Unit Bin



(c) 4 Objects in One Unit Bin

Curse of Dimensionality

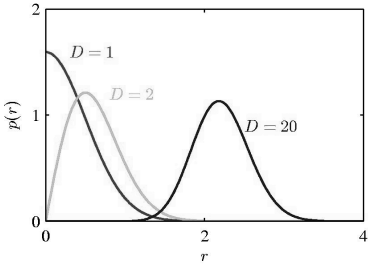
Polynomial curve fitting, M = 3

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} w_{ijk} x_i x_j x_k$$

Poly of order M,
no of coeff grows D^M

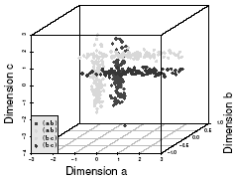
Gaussian Densities in higher dimensions

p(r): density
r: radius



Curse of Dimensionality

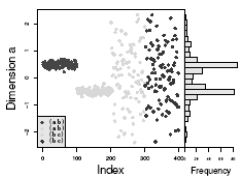
- **Curse of dimensionality does not prevent us from finding solutions applicable to high dimensional spaces. Because:**
- **i) Data is usually confined to a space having an effective smaller dimension (manifolds)**
- **ii) Data exhibit smoothness properties.**



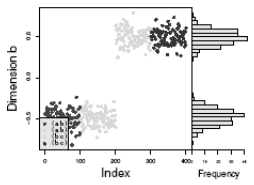
Why Subspace Clustering?

(adapted from Parsons et al. SIGKDD Explorations 2004)

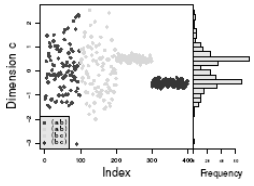
- **Clusters may exist only in some subspaces**
- **Subspace-clustering: find clusters in some of the subspaces**



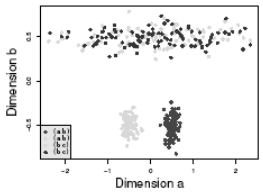
(a) Dimension a



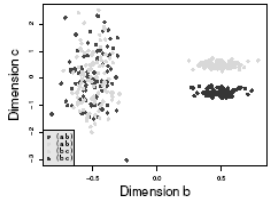
(b) Dimension b



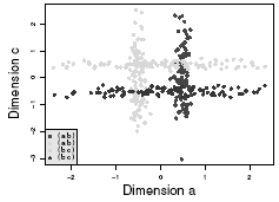
(c) Dimension c



(a) Dims a & b



(b) Dims b & c



(c) Dims a & c