Data Mining in R
Approccio non supervisionato
ANALISI CLUSTER
4. Validazione e interpretazione dei risultati

Laura Grassini

Tan, Steinbach, Kumar: Introduction to data mining, 2006, Addison Wesley
http://www-users.cs.umn.edu/~kumar/dmbook/index.php

Validazione

- Interpretazione soggettiva
- Misure oggettive di validazione
- Cluster e «very big data»

Intepretazione soggettiva

1) Capire se i gruppi sono fra loro diversi rispetto alle variabili di cluster

- Scomposizione della varianza delle variabili di cluster
- Box-plot per gruppo delle variabili di cluster

2) Interpretazione dei gruppi

Interpretiamo queste differenze. Possiamo analizzare anche eventuali variabili che non hanno partecipato all'analisi cluster, ma allora siamo in una validazione supervisionata.

Misure oggettive di validazione interna

Le misure di validazione interna si propongono di mostrare:

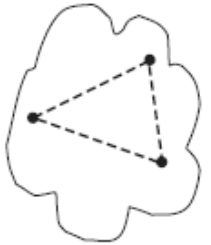
- **coesione** (cohesion): quanto i cluster sono omogenei al loro interno
- **separazione** (*separation*): l'entità della separazione fra cluster

Le misure dipendono anche dal tipo di algoritmo usato per la clusterizzazione.

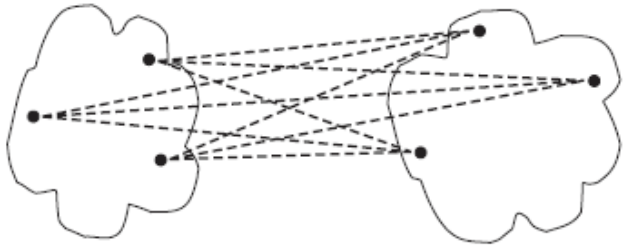
Servono per confrontare più soluzioni di clusterizzazione.

Si parla di **validazione interna** perché si usano gli stessi dati utilizzati per la clusterizzazione.

Coesione e separazione



(a) Cohesion.



(b) Separation.

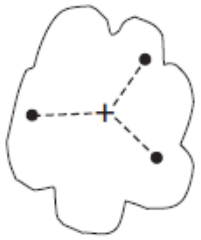
$$\text{Cohesion}(C_g) = \sum_{U_i, U_j \in C_g} distanza(U_i, U_j)$$

$$\text{Separation}(C_g, C_h) = \sum_{U_i \in C_g, U_j \in C_h} distanza(U_i, U_j)$$

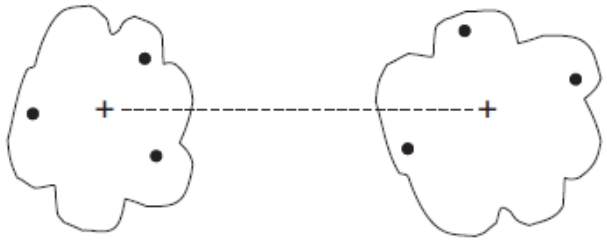
Per metodi *prototype based* (Ward, *k-means, PAM)*

$$\text{Cohesion}(C_g) = \sum_{U_i \in C_g} distanza(U_i, Prototype_g)$$

$$\text{Separation}(C_g, C_h) = distanza(Prototype_g, Prototype_h)$$

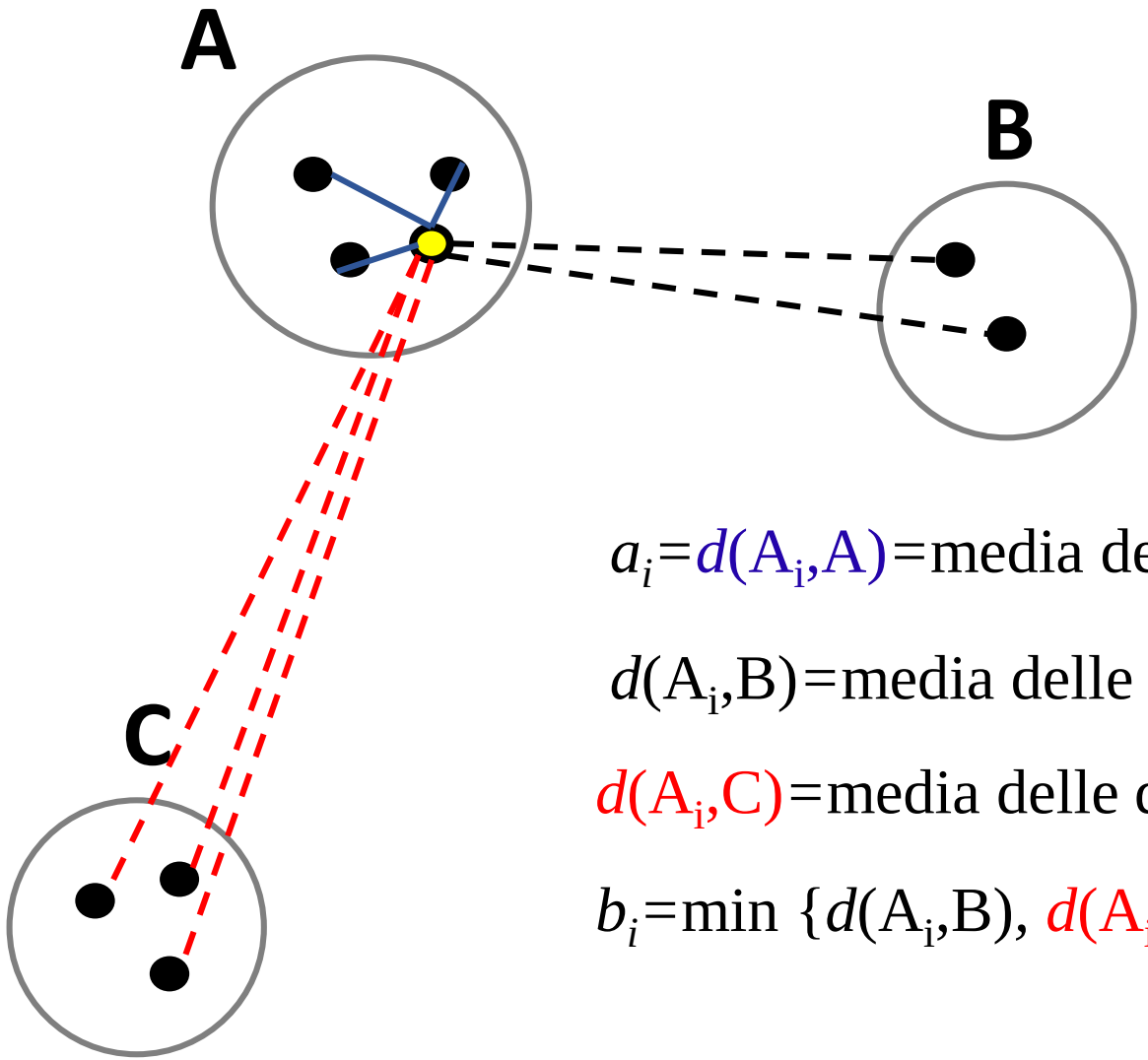


(a) Cohesion.



(b) Separation.

Silhouette (Rousseeuw, 1987)



A_i è il punto giallo nel cluster A

$$s_i = \frac{b_i - a_i}{\max(a_i; b_i)}$$

$a_i = d(A_i, A)$ = media delle distanza(*) fra A_i e gli altri punti in A

$d(A_i, B)$ = media delle distanza (*) fra A_i e i punti in B

$d(A_i, C)$ = media delle distanza (*) fra A_i e i punti in C

$b_i = \min \{d(A_i, B), d(A_i, C)\}$

(*) distanza o dissimilarità

Silhouette (continua)

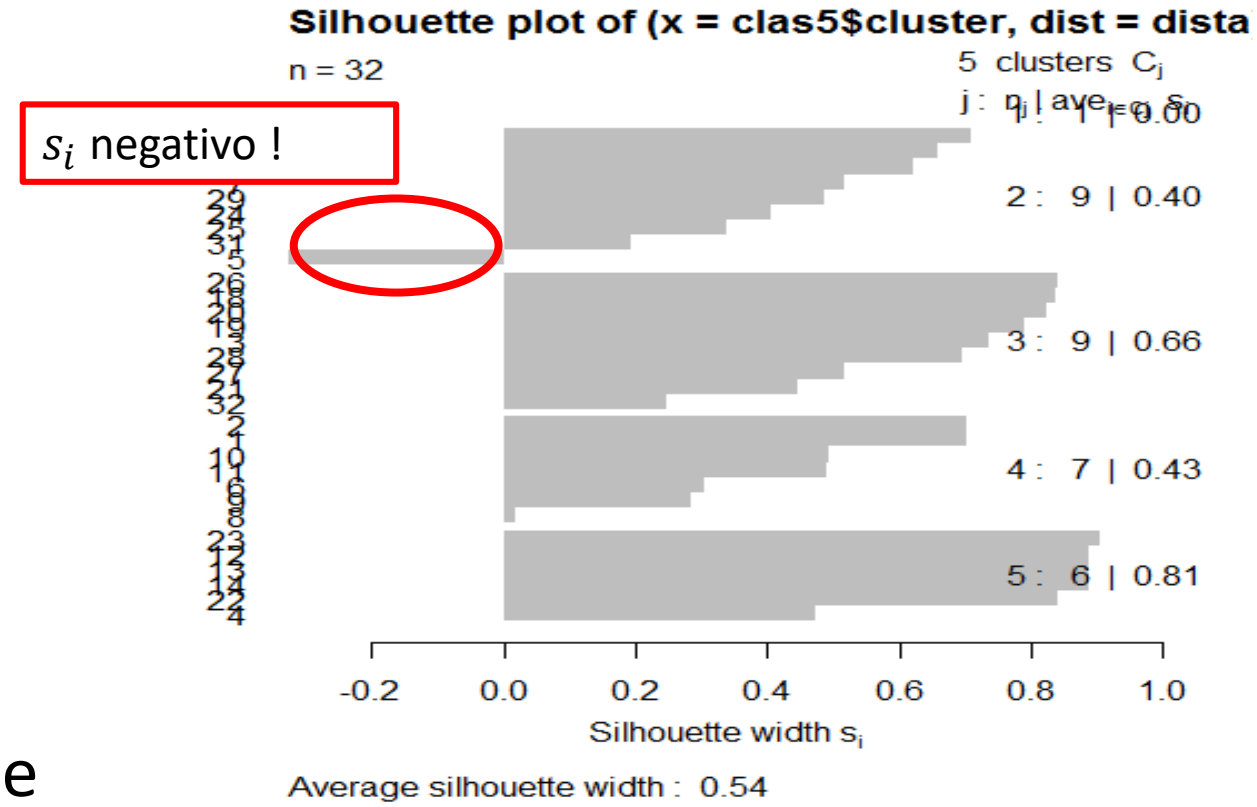
a_i=d(A_i,A)=media delle distanza fra A_i e gli altri punti in A

b_i=$\min_J$ {d(A_i,J)}, J≠A

$$s_i = \frac{b_i - a_i}{\max(a_i; b_i)}$$

$$\bar{s} = \frac{1}{n}\sum_{k=i}^{n} s_i$$

$\bar{s}$ è la valutazione della partizione ottenuta (valori più alti, risultati migliori)



Silhouette in R su `iris` dataset: selezione del n. di gruppi

```
### scelta del numero di gruppi
### kmeans e criterio silhouette
library(fpc)
## average silhouette: più grande è meglio
z<-iris[,1:4]
clusruns_asw<-kmeansruns(z, krange=2:5,iter.max=1000,
criterion='asw')
dista<-(dist(z))^2
plot(silhouette(clusruns_asw$cluster,dista))
```

Cluster e «very big data»