Statistica: Cenni sull'analisi dei gruppi
Clusters Analysis

Giovanna Jona Lasinio - Alessio Pollice (Università di Bari)

Dipartimento di Scienze Statistiche
Università di Roma "La Sapienza"

Ecobiologia e Scienze Naturali

Scopo dell'analisi dei gruppi

L'analisi dei gruppi mira ad assegnare le unità a categorie non definite a priori, formando dei gruppi di osservazioni omogenei al loro interno ed eterogenei tra loro. Cioè a costruire una *classificazione* delle osservazioni.

- Ricerca tipologica o individuazione di gruppi di unità con caratteristiche distintive;
- Stratificazione di popolazioni da sottoporre a campionamento;
- Definizione di sistemi di classificazione o tassonomie;
- Ricostruzione di valori mancanti tramite le informazioni desunte dal gruppo di appartenenza individuato tramite i dati disponibili;
- Sintesi delle osservazioni.

- Quasi tutte le tecniche considerano una *matrice di dissomiglianza* (somiglianza) che contiene le informazioni riguardanti il grado di dissomiglianza tra le diverse unità statistiche.
- La matrice di dissomiglianza può risultare da considerazioni soggettive sulle differenze tra le unità, come da calcoli effettuati sulla matrice dati.
- In questo secondo caso vi sono diversi criteri a seconda che le variabili rilevate siano, quantitative, qualitative, binarie o miste.

Matrice di dissomiglianza

$$\begin{pmatrix} 0 & d_{1,2} & \cdots & d_{1,n} \\ d_{1,2} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & d_{n-1,n} \\ d_{1,n} & \cdots & d_{n-1,n} & 0 \end{pmatrix}$$

Varabili Quantitative

La dissomiglianza tra unità viene valutata con una misura di distanza tra le stesse.
Indichiamo con $\mathbf{X}$ la matrice dei dati con n righe e k colonne, X_i indica la riga i-esima della matrice dei dati e x_{ih} indica l'elemento di $\mathbf{X}$ nella riga i colonna h. Vediamo ora alcuni tipi di distanze:

Varabili Quantitative

- *Distanza city block o di Manhattan*

$$d_{ij} = \sum_{h=1}^{k} |x_{ih} - x_{jh}|$$

- *Distanza euclidea*

$$d_{ij} = \sqrt{\sum_{h=1}^{k} |x_{ih} - x_{jh}|^2}$$

- *Distanza di Minkowsky*

$$d_{ij} = \sqrt[r]{\sum_{h=1}^{k} |x_{ih} - x_{jh}|^r}$$

Queste tre distanze sono fortemente influenzate dalle variabili (colonne di **X**) con varianza più alta, è quindi buona norma utilizzare **variabili standardizzate**

Variabili binarie

Assumiamo ora che gli elementi di $\mathbf{X}$ assumano solo valori $(0, 1)$, quindi prese due osservazioni X_i e X_j (righe) abbiamo

X_i	X_j	
	1	0
1	a	b
0	c	d

con $a + b + c + d = k$ ed

$$
\begin{aligned}
a &= \text{numero di variabili che valgono 1 per } X_i \text{ ed } X_j \\
b &= \text{numero di variabili che valgono 1 per } X_i \text{ e 0 per } X_j \\
c &= \text{numero di variabili che valgono 0 per } X_i \text{ e 1 per } X_j \\
d &= \text{numero di variabili che valgono 0 per } X_i \text{ ed } X_j
\end{aligned}
$$

Variabili binarie

- *Coefficiente di dissomiglianza semplice*: proporzione delle variabili che risultano discordanti

$$d_{ij} = \frac{b + c}{k}$$

- *Coefficiente di Jaccard*: indicato per variabili dicotomiche asimmetriche, che indicano la presenza di una data caratteristica

$$d_{ij} = \frac{b + c}{a + b + c}$$

Variabili Qualitative

Possiamo sfruttare i risultati ottenuti con le variabili binarie.

- costruiamo il coefficiente di dissomiglinaza semplice, indicando con c_{ij} la frazione delle variabili che assumono lo stesso valore per le unità i-esima e j-esima, si ha

$$d_{ij} = 1 - c_{ij}$$

Matrice dei dati con variabili miste

- Indice di Gower

$$d_{ij} = 1 - \frac{\sum_{h=1}^{k} s_{ijh}}{k}$$

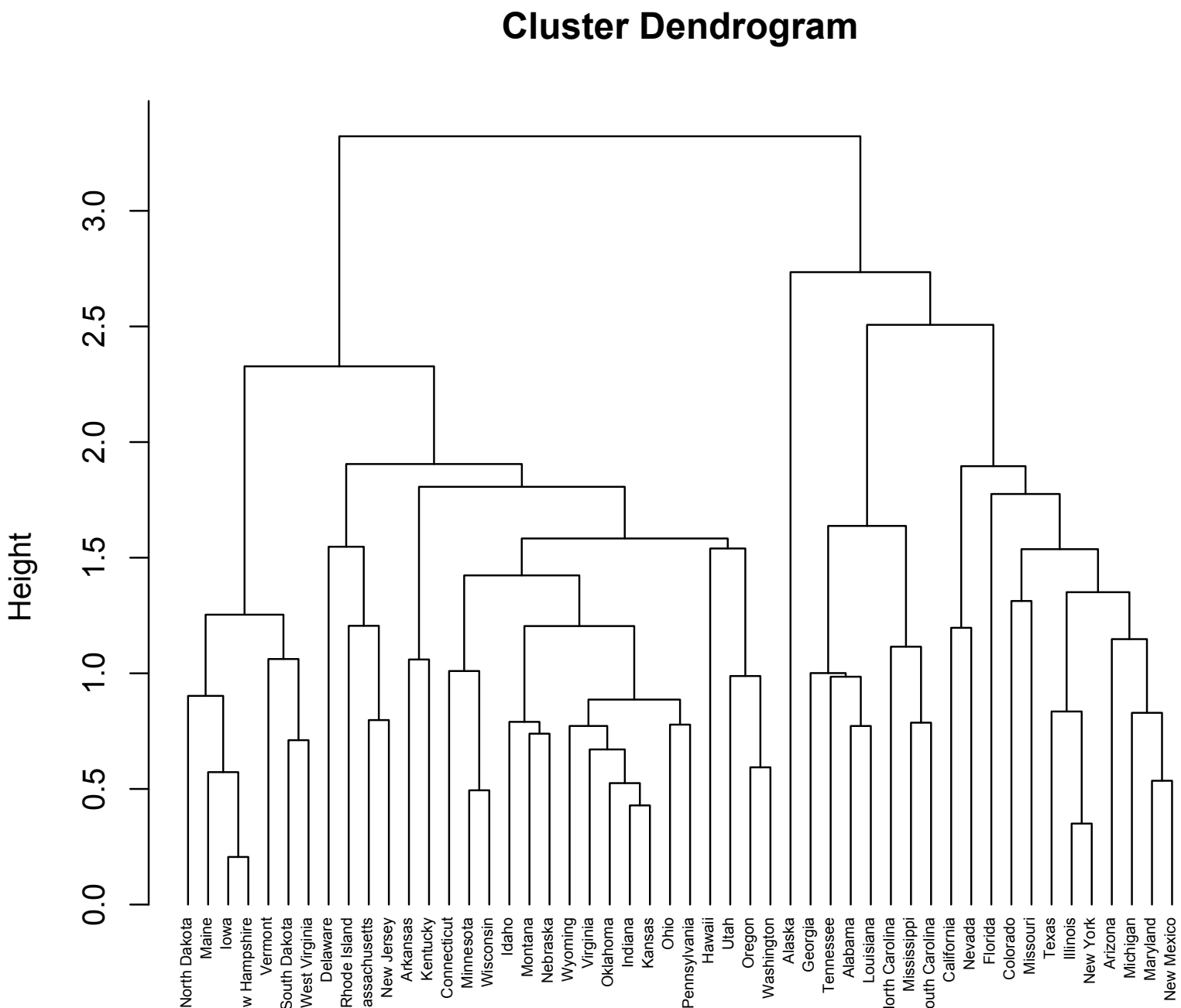
  - se la h-esima variabile è *quantitativa* ed $R(h)$ è il suo campo di variazione

$$s_{ijh} = 1 - \frac{|x_{ih} - x_{jh}|}{R(h)}$$

  - se la h-esima variabile è *qualitativa*

$$s_{ijh} = \begin{cases} 1 & \text{se ha la stessa modalità per le osservazioni } i\text{-esima e } j\text{-esima} \\ 0 & \text{altrimenti} \end{cases}$$

- Criteri per la creazione di partizioni annidate dell'insieme di osservazioni di partenza
  - Rappresentazione grafica della struttura di raggruppamento tramite *diagrammi ad albero* o *dendrogrammi*
  - Algoritmi *aggregativi* e *scissori*
  - Un oggetto allocato in un gruppo non può più venire riallocato in un diverso cluster in una fase successiva: un errore commesso nella fase iniziale della classificazione non può più essere messo in discussione


Cluster Dendrogram
Height
0.0
0.5
1.0
1.5
2.0
2.5
3.0
North Dakota
Maine
Iowa
w Hampshire
Vermont
South Dakota
West Virginia
Delaware
Rhode Island
assachusetts
New Jersey
Arkansas
Kentucky
Connecticut
Minnesota
Wisconsin
Idaho
Montana
Nebraska
Wyoming
Virginia
Oklahoma
Indiana
Kansas
Ohio
Pennsylvania
Hawaii
Utah
Oregon
Washington
Alaska
Georgia
Tennessee
Alabama
Louisiana
lorth Carolina
Mississippi
outh Carolina
California
Nevada
Florida
Colorado
Missouri
Texas
Illinois
New York
Arizona
Michigan
Maryland
New Mexico

- *Dendrogramma*:
  - Linee orizzontali rappresentano l'unione di due cluster
  - La loro posizione sull'asse delle ordinate indica la distanza alla quale i cluster vengono aggregati
  - Le ripartizioni in un certo numero di cluster possono essere ottenute sezionando opportunamente il dendrogramma
  - un criterio consiste nel sezionare il dendrogramma in corrispondenza del massimo scarto tra i livelli di prossimità ai quali avvengono le aggregazioni (ma le procedure gerarchiche non comprendono in genere alcun indicatore numerico che supporti la scelta del numero corretto di cluster)

- Dal collettivo non suddiviso si procede per *aggregazioni successive* generando gruppi sempre più numerosi, con un procedimento articolato in tre fasi:
  (i) *costruzione* della matrice di dissomiglianza in base alla misura prescelta
  (ii) *raggruppamento* degli elementi più somiglianti
  (iii) *calcolo* della matrice di dissomiglianza tra gruppi e/o singole osservazioni

  Le ultime due fasi vengono ripetute iterativamente finché tutti gli elementi non vengono aggregati in un unico cluster

Dissomiglianza tra cluster e/o singole osservazioni

- *legame singolo*: la distanza tra due gruppi equivale alla minore delle distanze tra gli elementi dei due gruppi
- *legame completo*: la distanza tra due gruppi equivale alla maggiore delle distanze tra gli elementi
- *legame medio*: la distanza tra due gruppi equivale alla media delle distanze tra gli elementi
- *centroide*: la distanza tra due gruppi equivale a quella tra i due centroidi corrispondenti
- *metodo di Ward*: la distanza tra due gruppi equivale alla devianza tra i centroidi

- Si parte dalla situazione in cui le n unit à fanno parte di un unico gruppo e si perviene alla formazione di n gruppi ognuno composto da una sola unit à
- Permettono la formazione di un numero qualsiasi di sottogruppi da un unico gruppo genitore
- *Metodo K-means*: inizia considerando la partizione in K gruppi del collettivo che minimizza la devianza interna ai gruppi e procede, ad ogni passo, suddividendo il gruppo avente devianza maggiore, in modo che la devianza interna complessiva risulti minima (equivale a massimizzare la distanza tra i centroidi dei gruppi)

- *Procedure iterative* che ammettono nelle diverse fasi una riallocazione degli elementi gi à clusterizzati, consentendo un progressivo miglioramento delle partizioni ottenute
- *Fasi*:
  - individuazione di una *partizione provvisoria* dei dati
  - ottimizzazione di una *funzione obiettivo*, modificando l'assegnazione

- *Metodo K-means*: tende a minimizzare la varianza interna ai gruppi individuati
  a. si determinano i *centroidi* della partizione iniziale
  b. si determina una *partizione provvisoria* attribuendo le osservazioni al centroide più vicino
  c. si calcolano i *nuovi centroidi* della partizione ottenuta
  d. si riprende da b.

  l'algoritmo si ferma quando due successive iterazioni conducono alla stessa partizione, oppure quando si raggiunge un numero di iterazioni prefissato, o ancora quando la funzione obiettivo non decresce più in modo significativo

Partitioning around medoids

Qui ci interessa introdurre due algoritmi di clustering di tipo *aggregativo*

- Partitioning Around Medoids
  - simile al metodo della k-medie ma molto più stabile ed affidabile
  - utilizza la PCA per costruire la visualizzazione dei clusters
  - fornisce misure utili per la valutazione della qualità dei gruppi ottenuti
- CLARA
  - funziona come PAM ma permette di trattare dataset molto grandi

Clusters: PAM

- Usa misure di dissimilarità invece delle distanze euclidee
- Definisce i gruppi attorno a centri che sono elementi del dataset su cui si lavora
- È dunque più stabile e robusto delle k-medie
- Fornisce misure di qualità dei gruppi ottenuti (silhouette)

Clusters: PAM, silhouette

- la misura di qualità viene costruita nel modo seguente:
  - Sia i un elemento qualsiasi del dataset e sia A il suo cluster di appartenenza
  - si calcola $a(i)$ ovvero la dissimilarità media (anche distanza euclidea) di i da tutti gli oggetti in A
  - si considera poi un qualunque cluster $C \neq A$ e si definisce $d(i, C)$ dissimilarità media tra i e gli elementi di C, si ripete questo per tutti i cluster $C \neq A$
  - si sceglie il valore più piccolo che si ottiene ($b(i)$), supponiamo sia relativo al cluster B

Clusters: PAM, silhouette

- il cluster B viene detto *vicino* di A
- quindi $s(i)$, il valore di silhouette per i viene definito come

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

- Appare evidente che $s(i) \in [-1, 1]$ quindi

$s(i) = 1$	$a(i) << b(i)$ quindi la dissimilarità tra i e gli elementi di A è bassa, i è molto ben classificata in A
$s(i) = 0$	$a(i) = b(i)$ i apparterrebbe egualmente bene sia ad A che a B
$s(i) < 0$	$a(i) < b(i)$ l'elemento è mal classificato, probabilmente va assegnato al cluster B

Clusters: PAM

- In output si ottengono due tipi di grafico utili
  - Cluster plot: Vengono calcolati i due primi assi principali e si su di essi si rappresentano i gruppi
  - Silhouette plot: Vengono rappresentati i valori di silhouette per ciascun elemento di un gruppo
- L'implementazione in R si trova nel pacchetto `cluster`, la funzione è `pam`

- I risultati di una strategia di raggruppamento dipendono oltreché dalla tecnica utilizzata, anche dal tipo di distanza e dall'uso di osservazioni grezze o standardizzate
- Rigidità delle strutture di raggruppamento annidate: un'aggregazione impropria effettuata nei primi passi dell'analisi viene portata avanti sino alla fine
- Le tecniche gerarchiche portano in genere a gruppi meno omogenei di quelli ottenibili attraverso le tecniche non gerarchiche

- Dal punto di vista del calcolo le tecniche gerarchiche risultano molto meno dispendiose di quelle non gerarchiche
- Dal punto di vista della quantità di memoria necessaria per effettuare l'analisi, mentre per le tecniche gerarchiche bisogna tenere in memoria almeno $n(n-1)/2$ numeri (quanti sono gli elementi della matrice di dissomiglianza), le tecniche non gerarchiche richiedono di tenere in memoria la sola matrice dati
- I risultati delle tecniche non gerarchiche sono modesti se non si dispone di una buona partizione iniziale
- Tra le tecniche gerarchiche aggregative la scelta deve essere effettuata nel rispetto della natura dei dati da analizzare