Cluster Analysis

Misure di distanza e dissimilarità

R. Pappadà (rpappada@units.it)

18 aprile 2024

Introduzione

Introduzione

- **Statistica descrittiva**
  insieme di tecniche per fornire una rappresentazione sintetica, grafica e numerica, delle rilevazioni di alcune caratteristiche relative a un insieme di unità che costituiscono la popolazione (o sotto-popolazione) oggetto di studio
- **metodi di analisi**
  differenti a seconda che le caratteristiche rilevate siano quantitative (cioè misurabili) o qualitative (attributi)
- **Dati**
  I risultati della rilevazione di una caratteristica (o una loro codifica) sono costituiti da una lista di numeri o codici. E' generalmente comodo usare codici numerici anche quando la caratteristica è di tipo qualitativo.

Statistica descrittiva multivariata

La statistica descrittiva multivariata ha lo scopo di descrivere ed estrarre informazioni relative a dati multidimensionali, ovvero unità sperimentali su cui sono state osservate più variabili (qualitative o quantitative)

Si dispone di un insieme di dati composto da n unità statistiche per le quali vengono osservate p variabili ($p > 2$) (che possono essere di vario tipo)

Alcune tecniche di analisi multivariata comunemente impiegate sono: le tecniche di raggruppamento o *cluster analysis*; l'analisi delle componenti principali; regressione lineare e sue estensioni

Cluster analysis

Lo scopo della cluster analysis è quello di raggruppare le unità statistiche in gruppi secondo un dato criterio di *similarità*, in modo che le osservazioni siano il più possibile omogenee all'interno dei gruppi ed il più possibile disomogenee tra gruppi diversi.

Esistono molti metodi di cluster analysis: una prima suddivisione è quella in algoritmi gerarchici e algoritmi non gerarchici (il più diffuso dei quali è il metodo delle K-medie)

Gli algoritmi di clustering sono tecniche di unsupervised learning, poiché hanno lo scopo di scoprire strutture presenti nei dati che non sono etichettati o classificati in determinate categorie

Misure di distanza e dissimilarità

Introduzione

Il concetto di dissimilarità tra due unità è centrale per molti metodi di *cluster analysis*:

- come passo preliminare, dobbiamo prima sviluppare una scala quantitativa su cui misurare la *similarità* tra gli oggetti
- la scelta della misura di (dis)similarità ha un forte impatto sui clusters risultanti
- occorre prestare particolare attenzione al tipo di dati a disposizione e alla domanda di ricerca

Esempio

Per un insieme di utenti di un servizio (unità) si dispone di alcune variabili come la spesa annuale per alcuni beni, la preferenza espressa verso alcuni prodotti, informazioni demografiche, ecc.
Si vuole quantificare il grado di "somiglianza"–similarità–o per coppie di utenti, sulla base di tutte le variabili disponibili

Spesso si misura la dissimilarità in termini di una misura di distanza

- presi due utenti, i ed r, per i quali disponiamo di p variabili $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, $\mathbf{x}_r = (x_{r1}, \ldots, x_{rp})$, definiamo alcune misure di dissimilarità/distanza più comuni
- due unità (utenti) 'simili' presenteranno un valore piccolo della distanza o grande in caso di coefficienti di similarità

Distanza: definizione

Consideriamo un insieme $\mathbf{X}$ di unità descritte da vettori p-dimensionali. Una distanza è una funzione che associa a ciascuna coppia di unità un numero reale

$$d : \mathbf{X} \times \mathbf{X} \to \mathbb{R}_+$$

tale che, per ogni scelta di unità $\mathbf{x}_i, \mathbf{x}_r, \mathbf{x}_s$ in $\mathbf{X}$, si abbia

(i) $d(\mathbf{x}_i, \mathbf{x}_r) \geq 0$ (non-negatività)

(ii) $d(\mathbf{x}_i, \mathbf{x}_r) = d(\mathbf{x}_r, \mathbf{x}_i)$ (simmetria);

(iii) $d(\mathbf{x}_i, \mathbf{x}_r) \leq d(\mathbf{x}_i, \mathbf{x}_s) + d(\mathbf{x}_s, \mathbf{x}_r)$ (disuguaglianza triangolare)

Inoltre $d(\mathbf{x}_i, \mathbf{x}_r) = 0$ se e solo se $\mathbf{x}_i = \mathbf{x}_r$.

Indici di dissimilarità

Una misura o indice di dissimilarità δ soddisfa le proprietà di

(i) $\delta(\mathbf{x}_i, \mathbf{x}_r) \geq 0$ (non-negatività) e $\delta(\mathbf{x}_i, \mathbf{x}_r) = 0$ se $\mathbf{x}_i = \mathbf{x}_r$.

(ii) $\delta(\mathbf{x}_i, \mathbf{x}_r) = \delta(\mathbf{x}_r, \mathbf{x}_i)$ (simmetria);

Una distanza è pertanto anche una dissimilarità ma non vale il viceversa.

Distanze per variabili quantitative

Distanza di Minkowsky tra $\mathbf{x}_i$ e $\mathbf{x}_r$

$$d_m(\mathbf{x}_i, \mathbf{x}_r) = \left\{ \sum_{k=1}^{p} |x_{ik} - x_{rk}|^m \right\}^{1/m} \qquad m \geq 1$$

- $\mathbf{m = 1} \rightarrow$ *Distanza di Manhattan* o *distanza "City Block"*

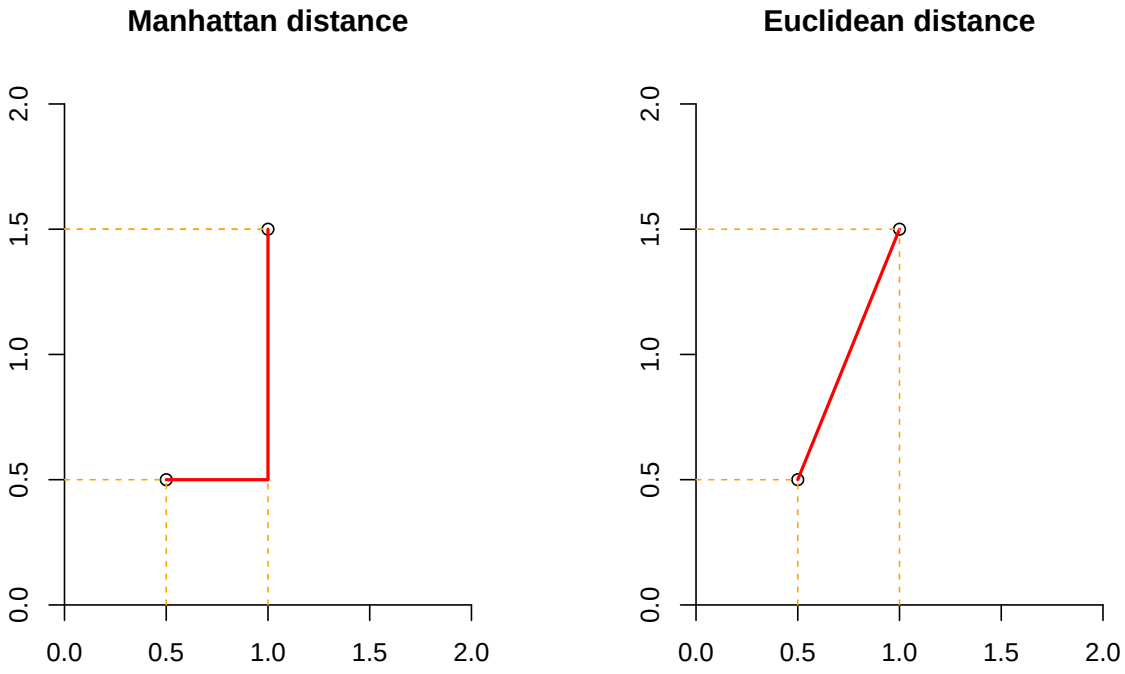
$$d_1(\mathbf{x}_i, \mathbf{x}_r) = \sum_{k=1}^{p} |x_{ik} - x_{rk}|$$

- $\mathbf{m = 2} \rightarrow$ *Distanza euclidea*

$$d_2(\mathbf{x}_i, \mathbf{x}_r) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{rk})^2}$$

La distanza di Minkowsky si applica a variabili numeriche.

Distanze per variabili quantitative



Distanza di Manhattan e distanza euclidea tra i punti $\mathbf{x}_i = (0.5, 0.5)$ *e* $\mathbf{x}_r = (1.0, 1.5)$.

Altre misure per variabili quantitative

Distanza del massimo

$$d(\mathbf{x}_i, \mathbf{x}_r) = \max_k |x_{ik} - x_{rk}|$$

Talvolta si usa la *distanza euclidea al quadrato*:

$$d_2^2(\mathbf{x}_i, \mathbf{x}_r) = \sum_{k=1}^{p} (x_{ik} - x_{rk})^2$$

Questa però non è una distanza secondo la definizione fornita prima, in quanto non vale la disuguaglianza triangolare!

Standardizzazione delle variabili

Il risultato di una analisi di clustering può essere influenzato dalla dispersione delle singole variabili: se una variabile ha una varianza molto alta, questo fa si che i gruppi siano determinati principalmente dai valori di questa variabile.

Per uniformare la dispersione delle variabili si ricorre alla standardizzazione dei dati, trasformando i dati in modo che ciascuna abbia peso uguale al reciproco della dispersione (più una variabile è dispersa meno peso le si attribuisce)

Standardizzazione delle variabili

Esempi

- variabili *centrate* e divise per la deviazione standard:

$$x^\star = \frac{x - \bar{x}}{sd(x)}$$

  Tutte le variabili assumono media nulla e varianza unitaria.

- normalizzazione in base al *range*: si divide ciascuna variabile per l'ampiezza dell'intervallo su cui assume valori la variabile stessa

$$x^\star = \frac{x - \min x}{\max(x) - \min(x)}$$

Indici di similarità per variabili binarie

Si assuma che tutte le variabili osservate su un insieme di unità siano binarie: possono assumere solo due valori che codificano, ad esempio la *assenza* o *presenza* di una caratteristica.

Ad esempio se si codificano i valori che denotano *assenza* con 0 e *presenza* con 1

	variabili				
	1	2	3	4	5
unità i	1	0	0	1	1
unità r	1	1	0	1	0

La *distanza euclidea al quadrato* tra le due unità è

$$d_2^2(\mathbf{x}_i, \mathbf{x}_r) = (1-1)^2 + (0-1)^2 + (0-0)^2 + (1-1)^2 + (1-0)^2 = 2$$

➔ $d_2^2 =$ n. valori differenti

Similarità per variabili binarie

Osservazione Viene data la stessa rilevanza ai seguenti due casi

- co-presenza di una caratteristica (1 & 1)
- co-assenza di una caratteristica (0 & 0)

Domanda: la co-assenza fornisce informazioni sulla similarità tra due unità?

A seconda della definizione di indice di similarità per dati binari possiamo gestire le co-occorrenze 0-0 in modo differente

- le coppie 0-0 hanno lo stesso peso delle coppie 1-1
- le coppie 1-1 hanno maggiore peso delle coppie 0-0, poiché la co-presenza è indicativa di similarità, mentre non è così per la co-assenza (es. informazione mancante)

Similarità per variabili binarie

Per ciascuna coppia, i valori osservati di p variabili binarie (0-1) possono essere organizzati come nella seguente tabella

$r \backslash i$	1	0	n. var
1	a	b	$a+b$
0	c	d	$c+d$
n. var.	$a+c$	$b+d$	$a+b+c+d=p$

- $a =$ numero co-presenze 1-1 per (i, r)
- $b =$ numero di caratteristiche presenti in r e assenti in i
- $c =$ numero di caratteristiche assenti in r e presenti in i
- $d =$ numero co-assenze 0-0 per (i, r)

Usando i dati dell'esempio precedente, si ha quindi $a = 2$, $b = c = d = 1$.

Similarità per variabili binarie

Per ogni coppia di unità, i ed r, possiamo definire un indice (o coefficiente) di similarità $s_{ir} \in [0, 1]$ tale che $s_{ir} = 1$ se le due unità presentano esattamente gli stessi valori per ogni variabile: $x_{ik} = x_{rk}$, per ogni $k = 1, 2, \ldots, p$.

(s1) Simple Matching	$s_{ir} = (a + d)/(a + b + c + d)$
(s2) Jaccard (Jaccard, 1908)	$s_{ir} = a/(a + b + c)$
(s3) Rogers and Tanimoto (1960)	$s_{ir} = (a + d)/(a + 2(b + c) + d)$
(s4) Sneath and Sokal (1973)	$s_{ir} = a/(a + 2(b + c))$

Tabella 1: Alcuni indici di similarità per variabili binarie

- Il coefficiente di Jaccard (s2) o il coefficiente (s4) ignorano la co-assenza (d) e quindi sono utili quando la co-assenza non è informativa
- un indice che invece è impiegato quando la co-assenza è rilevante è il coefficiente *simple matching*

Misure per dati binari: esempio

La tabella seguente riassume il comportamento dei primi due visitatori (2 delle n unità statistiche) di un sito di e-commerce nei confronti delle $p = 28$ pagine web (variabili) che essi possono visitare. Ogni pagina è considerata una variabile dicotomica: 1 =visitata; 0 =non visitata.

A \ B	1	0	tot
1	2	4	6
0	1	21	22
tot	3	25	$p = 28$

$a = 2$ (pagine visitate da entrambi gli utenti); $b = 4$, $c = 1$ (pagine visitate da uno ma non dall'altro); $d = 21$ (pagine non visitate)
Gli indici (s1) ed (s2) sono

- *simple matching* $s_{AB} = 23/28 \approx 0.821$
- *Jaccard* $s_{AB} = 2/(2 + 4 + 1) \approx 0.286$

Da similarità a dissimilarità

Un coefficiente di similarità s_{ij} può essere trasformato in un indice di dissimilarità δ_{ij} calcolando, per ogni $i, j \in \{1, \ldots, n\}$,

$$\delta_{ij} = 1 - s_{ij}$$

Dall'esempio precedente si ottiene

- $\delta^{SM}_{A,B} = 1 - 0.821 = 0.179$ (simple matching)
- $\delta^{J}_{A,B} = 1 - 286 = 0.714$ (Jaccard)

Dati qualitativi con più modalità

I possibili risultati di una variabile qualitativa sono anche detti modalità o livelli (più usato per le variabili ordinali).

Se i dati sono costituiti da p variabili con m modalità, allora per ciascuna coppia (i, r), un indice di similarità è ottenuto sommando il contributo della k-ma variabile

$$s_{ir} = \frac{1}{p} \sum_{k=1}^{p} s_{ir,k} \qquad s_{ir} \in [0, 1]$$

dove

$$s_{ir,k} = \begin{cases} 1 \text{ se } x_{ik} = x_{rk} \\ 0 \text{ altrimenti} \end{cases}$$

Indici di similarità per dati misti

Esistono alcuni approcci per calcolare la dissimilarità per dati di cui si dispone di variabili quantitative, ordinali, binarie

L'indice di Gower è una misura di dissimilarità che aggrega il contributo derivante da ciascuna variabile

$$d_G(\mathbf{x}_i, \mathbf{x}_r) = \frac{\sum_{k=1}^{p} w_{ir,k}\delta_{ir,k}}{\sum_{k=1}^{p} w_{ir,k}},$$

dove generalmente si pone il peso $w_{ir,k} = 1$ (ad eccezione del caso in cui il dato sia mancante o quando si vuole attribuire un peso diverso ad una o a un gruppo di variabili), e $\delta_{ir,k}$ è la dissimilarità relativa alla k-ma variabile per (i, r)

Indice di Gower

a. Per variabili binarie e qualitative con più modalità

$$\delta_{ir,k} = \begin{cases} 0 \text{ if } x_{ik} = x_{rk} \\ 1 \text{ otherwise} \end{cases}$$

b. per variabili quantitative, posto $R_k = max(x_{ik}) - min(x_{rk})$,

$$\delta_{ir,k} = \frac{|x_{ik} - x_{rk}|}{R_k}$$

c. per variabili ordinali con m livelli, queste vengono codificate con i ranghi $x' \in \{1, \ldots, m\}$ e trasformate in $x'' = \frac{x' - 1}{m - 1} \in [0, 1]$; per la dissimilarità si utilizza [b.]

Matrice di distanza/dissimilarità

Le distanze/dissimilarità fra n unità possono essere rappresentate tramite una matrice quadrata di dimensione n, simmetrica, con 0 sulla diagonale principale:

$$\Delta = \begin{bmatrix} 0 & \delta_{12} & \dots & \delta_{1j} & \dots & \delta_{1n} \\ \delta_{21} & 0 & \dots & \delta_{2j} & \dots & \delta_{2n} \\ \vdots & & & \vdots & & \vdots \\ \delta_{j1} & \delta_{j2} & \dots & 0 & \dots & \delta_{jn} \\ \vdots & & & \vdots & & \vdots \\ \delta_{n1} & \delta_{n2} & \dots & \delta_{nj} & \dots & 0 \end{bmatrix}$$

Commenti finali

La matrice di dissimilarità o distanza è utilizzata come input per molti algoritmi di cluster analysis.

I risultati di una strategia di raggruppamento dipendono oltre che dalla tecnica utilizzata, anche dal tipo di distanza e dall'uso di osservazioni grezze o standardizzate.

In generale, una buona tecnica di raggruppamento dovrebbe garantire risultati identici se applicata sugli stessi dati da persone diverse indipendentemente (*oggettività*), la *stabilità* dei risultati a piccole variazioni nei dati, l'*informatività* del raggruppamento finale e la *semplicità* dal punto di vista algoritmico.