Analisi Statistica per le Imprese

Prof. L. Neri

Dip. di Economia Politica e Statistica

Cap 5.4 Analisi dei gruppi

Premessa

Consideriamo un certo numero n di unità su cui abbiamo osservato p fenomeni (variabili).

- Obiettivo: Individuare gruppi di osservazioni all'interno dei quali le unità siano simili (omogenee al loro interno) ed eterogenee tra di loro (gruppi distinti).
- Tale omogeneità/disomogeneità si riferisce all'insieme delle variabili osservate
- Attenzione: non sappiamo a priori se tali gruppi esistono effettivamente

Attraverso l'AG, una realtà molto variegata viene semplificata e ricondotta ad alcune tipologie più leggibili: CLASSIFICAZIONE. L'obiettivo è quello di realizzare un raggruppamento rispetto a p caratteristiche, ma abbiamo bisogno di un algoritmo non banale (per p>3 la rappresentazione grafica non ci aiuta)

Impieghi di AG in ambito economico-aziendale

Identificazione di:

- gruppi di consumatori (o utenti di un certo servizio pubblico) sulla base di: comportamento al consumo, opinioni sul prodotto, importanza assegnata a varie caratteristiche di un prodotto (segmentazione del mercato)
- gruppi di strutture secondo varie caratteristiche che ne definiscono l'efficienza
- opportunità per potenziali nuovi prodotti
- mercati (aree test di mercato)
- aziende secondo specifiche caratteristiche

Fasi dell'AG

- Scelta delle variabili ed eventuale trasformazione delle stesse
- Scelta della misura di dissomiglianza
- Scelta dell'algoritmo di raggruppamento
- Valutazione della partizione ottenuta e scelta del numero ottimale di gruppi
- Interpretazione dei risultati ottenuti (connotazione dei gruppi)

Scelta delle variabili

- Una volta formulato il problema si devono scegliere le variabili alla base della procedura di raggruppamento. La metodologia statistica è di scarso aiuto, è necessaria una buona conoscenza del fenomeno (l'impiego di variabili con scarso potere discriminatorio può rendere confusa la classificazione).
- Le variabili selezionate dovrebbero essere dei buoni indicatori delle "similarità" tra le unità (similarità rilevanti per il problema oggetto di studio). Per eliminare in modo mirato la ridondanza di variabili si può applicare una tecnica di riduzione dei dati tipo *Analisi Fattoriale* o *Analisi delle Componenti Principali* e poi applicare l' AG sui punteggi ottenuti.
- Se le variabili selezionate sono variabili quantitative ed espresse secondo diverse unità di misura/diverso ordine di grandezza: *standardizzazione*

Scelta della misura di dissomiglianza

- Variabili quantitative: indice di distanza
- Variabili qualitative: indice di dissimilarità

Esistono molte funzioni di distanza e di dissimilarità, in generale per la scelta ci si basa sulle caratteristiche/proprietà delle singole metriche

- per variabili quantitative, la più usata è la distanza euclidea
- per variabili quantitative tra loro correlate, la più usata è la distanza di Mahalanobis

Osservazione: può essere opportuno verificare la stabilità dei risultati con vari tipi di distanza/dissimilarità.

Scelta della misura di dissomiglianza

Di seguito mostriamo in dettaglio solo alcune misure di dissomiglianza per caratteri quantitativi e qualitativi dicotomici

Similarità e distanza per caratteri quantitativi

Una misura di distanza deve godere delle seguenti proprietà:

1. $d_{ij} >= 0$ (non negatività)
2. $d_{ii} = 0$
3. $d_{ij} = d_{ji}$ (simmetria)
4. $d_{ij} <= d_{ir} + d_{rj}$ (diseguaglianza triangolare: la distanza che intercorre tra due punti e' sempre minore della somma delle distanze tra tali punti e un terzo punto)

Se una misura di distanza soddisfa tutte le quattro proprietà, si dice che lo spazio di riferimento è metrico

Distanza Euclidea

Siano x_i e x_j due vettori contenenti il profilo di due unità, misurato su p attributi. La distanza euclidea è definita dalla norma della differenza tra i vettori rappresentativi delle unità ovvero:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

Questa distanza corrisponde alla usuale distanza tra punti nello spazio fisico. Si osservi invece che facendone uso in campo statistico essa combina scarti tra grandezze che possono essere espresse in unita' di misura diverse. La somma non ha quindi nessun signicato a meno che le unita' di misura siano le stesse.

Misure di distanza per variabili qualitative dicotomiche

Supponiamo che la matrice X contenga p misurazioni nominali effettuate su n individui; in particolare, si valuta la presenza (1) o l'assenza (0) di p attributi

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Profilo unità i	1	1	0	1	1	0	1	0	0
Profilo unità j	1	0	1	1	0	1	1	1	0

Misure di distanza per variabili qualitative dicotomiche

Con riferimento alle due unità, possiamo sintetizzare le due righe della matrice dei dati mediante la seguente tabella di contingenza

		unità i	
		1	0
unità j	1	a	b
	0	c	d

dove

a rappresenta il numero dei caratteri presenti in entrambe le unità

b il numero dei caratteri presenti nell'unità j, ma assenti nell'unità i

c il numero dei caratteri presenti nell'unità i, ma assenti nell'unità j

d il numero dei caratteri assenti in entrambe le unità

Coefficiente di similarità di Jaccard

In letteratura sono presenti diversi modi di calcolare la similarità che differiscono principalmente per il trattamento riservato all'aggregato d. In particolare, trattiamo al similarità di Jaccard, per il quale si definisce

$$c_{ij} = a/(a+b+c)$$

il coefficiente di distanza è dato da

$$d_{ij} = 1 - c_{ij} = b+c/(a+b+c)$$

Misure di distanza per misurazioni ordinali

Una modo consiste nell'attribuire un punteggio alle categorie ed utilizzare una delle misure di distanza o similarità introdotte per i caratteri quantitativi. L'operazione contiene ovvi elementi di arbitrarietà.

Scelta dell'algoritmo di raggruppamento

- *I metodi gerarchici* consentono di ottenere un insieme di gruppi ordinabili secondo livelli crescenti, con un numero di gruppi da n ad 1:
  - al livello iniziale ogni unità costituisce un gruppo
  - negli stadi intermedi si aggregano gli elementi in gruppi via via sempre più numerosi
  - al livello finale tutte le unità sono riunite in un unico gruppo
  - la scelta del numero dei gruppi avviene contestualmente
- *I metodi non gerarchici* forniscono un'unica partizione delle n unità in g gruppi, g deve essere specificato a priori

Se si ritiene che nei dati vi sia una struttura gerarchica allora alg. gerarchici, altrimenti non gerarchico

Valutazione della partizione ottenuta

- Negli algoritmi di tipo gerarchico avviene, sostanzialmente, in base al seguente principio:
  - non bisogna accorpare gruppi troppo diversi tra loro.
- Negli algoritmi di tipo non gerarchico le soluzioni con g o $(g-1)$ gruppi sono confrontabili solo attraverso indici sintetici

Attenzione: l'esistenza dei gruppi determinati non è scontata, potremmo aver ottenuto una partizione che non esiste nella realtà.
Ci chiediamo quindi: la classificazione ottenuta fornisce gruppi

- composti da unità simili?
- distinti tra loro?

Interpretazione della partizione ottenuta

- Quali sono le caratteristiche dei singoli gruppi ottenuti?
- Oppure quali sono le differenze tra i gruppi?

E' evidente che nello svolgimento dell'analisi si introducono elementi di soggettività: è importante quindi verificare la stabilità della soluzione.

I metodi gerarchici: introduzione

- Sono così denominati perchè procedono per aggregazioni successive delle n unità statistiche.
- La procedura inizia con n gruppi, formati ciascuno da una unità, per poi passare a $(n-1)$ gruppi, $(n-2)$....fino ad arrivare a quella in cui tutte le unità sono riunite in 1 solo gruppo.
- Ovviamente l'obiettivo non è quello di arrivare ad un unico gruppo bensì quello di raggruppare le n unità in un numero g limitato di gruppi.

N.B.: una partizione in g gruppi sarà caratterizzata da una maggiore omogeneità interna rispetto alla partizione in $g-1$ gruppi.

Fasi dello schema di raggruppamento gerarchico

1. A partire dagli n gruppi, si calcola la matrice delle distanze D (simmetrica e $n \times n$).
2. Si individuano in D le due unità più simili (con minore distanza) e si riuniscono in un unico gruppo, formando quindi $(n-1)$ gruppi.
3. Si calcola una nuova matrice di distanza tra gruppi $(n-1 \times n-1)$, come? Poichè le unità oggetto di fusione non esistono più come soggetti singoli vengono eliminate dalla matrice delle distanze le due righe e le due colonne corrispondenti, ottenendo una nuova matrice delle distanze di dimensione $(n-2 \times n-2)$.
4. Si aggiunge una nuova riga ed una nuova colonna che contiene le distanze tra il nuovo gruppo (ottenuto dalla fusione precedente) e tutte le unità che continuano ad esistere singolarmente. Si ottiene una nuova matrice di dimensione $(n-1 \times n-1)$.

Matrice delle distanze

Le $n(n-1)/2$ distanze vengono raccolte nella matrice simmetrica:

$$D = \begin{bmatrix} 0 & d_{12} & & & d_{1j} & \\ & 0 & & & & \\ & & 0 & & d_{ij} & \\ & & & 0 & & \\ & & & & 0 & d_{n-1,n} \\ & & & & & 0 \end{bmatrix}$$

facendo riferimento all'esempio fatto in precedenza (reddito disponibile-consumo pro-capite per i comuni), consideriamo il reddito, abbiamo che la matrice delle distanze (calcolate come differenze dei termini prese in valore assoluto) è data da:

$$D = \begin{bmatrix} & A & B & C & D & E \\ A & 0 & 0.8 & 2.7 & 4.0 & 2.3 \\ B & & 0 & 3.5 & 3.2 & 1.5 \\ C & & & 0 & 6.7 & 5.0 \\ D & & & & 0 & 1.7 \\ E & & & & & 0 \end{bmatrix}$$

Definizione della funzione distanza

La definizione della distanza tra due gruppi definisce il metodo di raggruppamento:

- del legame singolo
- del legame completo
- del legame medio
- del centroide
- di Ward

Non vedremo nel dettaglio i metodi di raggruppamento, ci soffermiamo su due metodi a titolo esemplificativo

Metodo del legame singolo

La distanza tra gruppi e misurata dalla distanza più piccola esistente tra gli elementi appartenenti ad un gruppo e quelli appartenenti ad un altro

Consideriamo l'esempio precedente. La distanza minore è pari a $d_{12} = d_{21}$ quindi il comune A e il comune B vengono aggregati nel gruppo (AB) dopo aver ottenuto il cluster (AB) si considera la minore tra le distanze di A e B da tutti i restanti comuni ovvero

$$d_{ij(k)} = min\left(d_{ik}, d_{jk}\right)$$

tra le due distanze di volta in volta considerate si sceglie la minore.

Metodo del legame singolo

nell'esempio abbiamo

$$d_{(AB)C} = min(d_{AC}, d_{BC}) = d_{AC} = 2.7$$

$$d_{(AB)D} = min(d_{AD}, d_{BD}) = d_{BD} = 3.2$$

$$d_{(AB)E} = min(d_{AE}, d_{BE}) = d_{BE} = 1.5$$

e troviamo una nuova matrice delle distanze

$$D_1 = \begin{bmatrix} & (AB) & C & D & E \\ (AB) & 0 & 2.7 & 3.2 & 1.5 \\ C & & 0 & 6.7 & 5.0 \\ D & & & 0 & 1.7 \\ E & & & & 0 \end{bmatrix}$$

alla seconda iterazione è E che è il più vicino a (AB) alla distanza 1.5

Metodo del legame singolo

Calcolando le nuove distanze tra il cluster (ABE) e i restanti comuni C e D risulta

$$d_{(ABE)C} = min(d_{AC}, d_{BC}, d_{EC}) = d_{AC} = 2.7$$

$$d_{(ABE)D} = min(d_{AD}, d_{BD}, d_{ED}) = d_{AC} = 1.7$$

si perviene quindi alla nuova matrice delle distanze

$$D_2 = \begin{bmatrix} & (ABE) & C & D \\ (ABE) & 0 & 2.7 & 3.2 \\ C & & 0 & 6.7 \\ D & & & 0 \end{bmatrix}$$

alla terza iterazione si unisce D al livello di distanza 1.7 formando il cluster (ABED).
all'ultimo passo si calcola la distanza tra i due cluster rimasti

Metodo del legame singolo

L'ultima iterazione aggrega i due gruppi in un unico gruppo contenente tutte le unità. La sequenza delle fusioni è pertanto rappresentata nella tabella seguente:

Iterazione	Gruppi	Livello di distanza
0	(A)(B)(C)(D)(E)	-
1	(AB)(C)(D)(E)	0.8
2	(ABE)(C)(D)	1.5
3	(ABED)(C)	1.7
4	(ABEDC)	2.7

Metodo di Ward

- Tale metodo può essere impiegato con variabili quantitative e con qualsiasi distanza calcolabile per tale tipo di variabili.
- Per semplicità, sarà, tuttavia, presentato impiegando la distanza euclidea quadrato.

Ricordiamo che la Devianza totale delle p variabili è la somma delle distanze euclidee al quadrato tra le singole osservazioni ed il vettore delle medie :

$$T = \sum_{s=1}^{p} \sum_{i=1}^{n} (x_{iS} - \bar{x}_S)^2 = \sum_{i=1}^{n} \sum_{s=1}^{p} (x_{iS} - \bar{x}_S)^2 = \sum_{i=1}^{n} d^2(i, \bar{x})$$

dove $\bar{x}_S$ è la media della variabile s con riferimento all'intero collettivo. Data una partizione in g gruppi, tale devianza può essere scomposta in Devianza entro i gruppi e Devianza tra i gruppi:

$$Devtot(p) = Deventro(p) + Devtra(p)$$

$$T = W + B$$

Metodo di Ward

- Nel passare da $k+1$ a k gruppi (aggregazione): la Devianza entro aumenta, la Devianza tra diminuisce.
- per $k = g$ (primo passo) $Dev entro = 0$
- per $k = 1$(ultimo passo) $Dev entro = Dev tot$ e $Dev tra = 0$.

Ad ogni passo della procedura di Ward si aggregano tra loro quei gruppi per cui vi è il minor incremento della devianza entro i gruppi

Scelta del num. di gruppi e valutazione della partizione

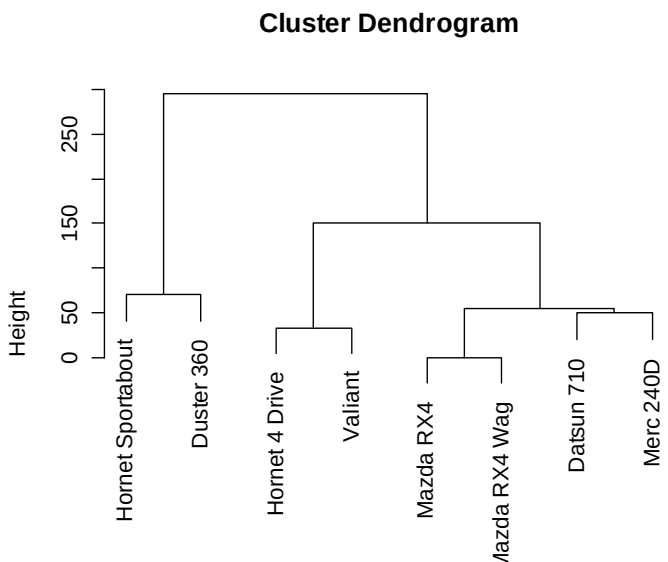
- In generale il criterio che si usa per la scelta del numero dei gruppi è il seguente:
  - si considerino due passi consecutivi nella procedura di aggregazione;
  - se nel passare da $k+1$ a k gruppi si aggregano due gruppi molto diversi tra loro, allora è meglio fermarsi prima, cioè a $k+1$ gruppi.

A tale fine possiamo impiegare varie tecniche:

- dendrogramma,
- scree plot (argomento non trattato)
- indice R quadro (argomento non trattato)

Il dendrogramma

- Un grafico che riporta sull'asse orizzontale, non quantitativo, le unità, e sull'asse verticale il livello distanza a cui avviene la fusione tra i diversi gruppi che si formano per aggregazioni successive.
- Dall'esame del dendrogramma si può ottenere un'importante indicazione in merito al numero di gruppi, in particolare il "taglio" avviene allo stadio in cui la distanza di fusione risulta troppo elevata



Il dendrogramma

- N.B. Aggregazioni che nel dendrogramma avvengono molto in alto si riferiscono a gruppi non omogenei tra loro. Quindi, se per esempio alle unità che compongono tale gruppo si va ad applicare un'unica strategia di marketing, verosimilmente si otterranno risposte eterogenee, probabilmente vanificando parte degli obiettivi dell'analisi
- E' dunque opportuno fermare il processo di aggregazione ad un livello di distanza minore, identificando un numero più elevato di gruppi, caratterizzati da maggiore omogeneità interna, e mettere a punto sui gruppi diverse strategie di marketing.

I metodi non gerarchici: introduzione

Gli algoritmi di tipo non gerarchico mirano a classificare le n unità statistiche in un numero prefissato di gruppi, specificato a priori. Si cerca la partizione in gruppi che soddisfi un determinato criterio di ottimalità attraverso:

- procedura iterativa in cui si definisce una partizione iniziale e si spostano successivamente le unità da un gruppo all'altro così da ottenere la partizione "ottimale".
- In genere "ottimale" corrisponde ad un criterio di minimizzazione della $Dev entro$.

Vantaggi:

- velocità di esecuzione
- non c'è più il vincolo per cui negli alg. gerarchici se due unità vengono fuse all'inizio, rimangono tali fino alla fine
- non necessita dell'uso del dendrogramma che, per n elevato, risulta difficilmente interpretabile

Fasi di una procedura iterativa alla base degli alg. di tipo non gerarchico

1. Scelta del numero g di gruppi
2. Scelta della classificazione iniziale in g gruppi (classificazione provvisoria)
3. Calcolo del valore della funzione obiettivo
4. Riallocazione delle unità in gruppi che garantiscono il miglioramento più elevato nella coesione interna ai gruppi
5. Iterazione dei passi 3 e 4 fino a che non viene soddisfatta una regola di arresto

Il metodo delle *k*-means

Il metodo più utilizzato nell'ambito dei metodi non gerarchici è quello delle *k-means*, in modo non rigoroso descriviamo i passi che esegue per allocare le unità ai g gruppi.

1. Scelta di g centri (poli, semi: $c_1, c_2, \ldots, c_h, \ldots, c_g$) nello spazio $p-dimensionale$
2. Raggruppamento delle unità intorno ai g centri in modo che il gruppo delle unità associate al centro c_h è costituita dall'insieme delle unità più vicine a c_h che a qualsiasi altro centro.
3. Calcolo dei centroidi dei g gruppi così ottenuti
4. Calcolo della distanza di ogni elemento da ogni centroide: se la distanza minima non è ottenuta in corrispondenza del centroide del gruppo di appartenenza, allora l'unità è riallocata al gruppo che corrisponde al centroide più vicino
5. Ricalcolo dei centroidi
6. Iterazione dei passi 4. e 5. fino a che non si raggiunge una configurazione stabile (tutte le unità vengono riassegnate allo stesso gruppo del passo precedente).

Il metodo k-means

- Solitamente si utilizza la distanza euclidea, che garantisce la convergenza della procedura iterativa.
- Il criterio di ottimalità corrisponde alla minimizzazione della $Dev entro$ i gruppi (W).
- un indice per la bontà della partizione è l'indice Calinski-Harabasz

Difetti:

- La classificazione finale può essere influenzata dalla scelta iniziale dei poli
- Soluzioni instabili se: vi sono valori anomali, se nei dati non esiste struttura in gruppi, se n piccolo
- Soluzione: meglio scelta casuale dei poli iniziali (badando a che i centri non siano valori anomali e che siano ben distinti) oppure scegliamo come centro il baricentro di una nube di punti.

Indice di Calinski-Harabasz

Per ogni numero di clusters $k \geq 2$l'indice di Calinski-Harabasz è dato da

$$ch(K) = \frac{devtra/(K-1)}{deventro/(n-K)}$$

il numero di clusters ottimo si ottiene in corrispondenza del k per cui vi sono grandi dissimilarità tra i clusters e grandi similarità dentro i clusters, la soluzione è quindi quel k che massimizza l'indice

Metodi gerarchici o non: conclusioni

- La scelta del metodo richiede una valutazione sulla base della tipologia dei dati.
- Se si dispone di variabili qualitative o miste, alcuni metodi risultano inapplicabili.
- Quando la matrice dei dati è costituita da sole variabili quantitative, la gamma dei metodi a disposizione è più ampia, in questo caso però è opportuno valutare l'opportunità di trasformazioni di variabili (standardizzazione) e anche valutare la struttura di correlazione tra le variabili (è inopportuno considerare più variabili fortemente correlate).
- In generale i metodi gerarchici risultano più sensibili ai valori anomali e non consentono di modificare un'aggregazione fatta al passo precedente.

Metodi gerarchici o non: conclusioni

- Una strategia spesso utilizzata consiste nel ricorrere prima ad un'analisi gerarchica per individuare il numero ottimale dei gruppi ed i punti iniziali (definiti come medie di gruppo), da assegnare in input ad un algoritmo non gerarchico con il quale definire la soluzione finale.
- Una strategia alternativa può essere quella di effettuare un'analisi non gerarchica, assegnando un numero iniziale elevato di gruppi al fine di individuare sia valori anomali, che la procedura dovrebbe individuare creando gruppi di una sola unità, sia i gruppi più significativi, dalle cui medie ricavare i punti iniziali da dare in input per una nuova analisi non gerarchica.

 L. Molteni, G. Troilo, 2007, Ricerche di Marketing, seconda edizione McGraw-Hill

 Bracalente B., Cossignani M., Mulas A., 2009, Statistica Aziendale, McGraw-Hill.