10.1 Il problema della classificazione

Classificare significa – riportando le parole di Italo Scardovi – "aggruppare per somiglianze e differenze rispetto a più caratteri, sostituendo alla vaga pluralità degli enti singoli la gradualità tipologica delle classi" [Mignani e Montanari 1994].

Il problema della *classificazione* si configura come l'individuazione di metodologie che consentano di cogliere la presenza di *gruppi* di unità statistiche omogenei al loro interno e tra loro separati rispetto a un determinato insieme di variabili. Mentre nell'ambito delle scienze fisiche, biologiche e naturali i gruppi pre-esistono operativamente e concettualmente al processo di classificazione e vanno solamente individuati (e non definiti), nell'ambito delle scienze sociali essi sono spesso il prodotto diretto della classificazione adottata, vengono cioè definiti proprio tramite essa.

Qualunque sia il contesto in cui si opera, riveste una fondamentale importanza la scelta delle variabili da coinvolgere nella classificazione, cioè la scelta delle variabili in base alle quali si vogliono individuare somiglianze e differenze tra le singole unità statistiche. Da un lato la presenza di variabili ugualmente distribuite nei diversi gruppi può nascondere, o comunque rendere di più difficile individuazione, la presenza dei gruppi stessi; dall'altro, l'uso di variabili che differenziano fortemente i vari gruppi ma che non hanno saldi legami con i fini della ricerca può portare all'individuazione di gruppi privi di significato, e nello stesso tempo l'esclusione di variabili con un elevato potere discriminante può togliere validità alla ricerca effettuata. La possibilità di fare riferimento a una disponibilità anche molto elevata di informazioni sulle unità statistiche coinvolte deve

scendere a patti con l'esigenza di ottenere una classificazione mirata.

Storicamente il concetto di classificazione è piuttosto "antico", ma il problema è stato affrontato per la prima volta in un'ottica statistica solamente sul finire del diciannovesimo secolo, con gli studi di Karl Pearson.

Le teorie della classificazione hanno dato luogo a numerose e diversificate pubblicazioni specialistiche. Il primo lavoro importante di riferimento è stato quello di Sokal e Sneath dei primi anni Sessanta; sono poi seguiti i contributi di Lance e Williams, Lerman, Anderberg, Benzécri, Hartigan, Lerman, Gordon.

In questo Capitolo verranno illustrate le tecniche di *analisi dei gruppi* più diffuse e utilizzate [→ Lebart, Morineau e Piron 1995; Mignani e Montanari 1994; Monari 1997].

10.2 Le tecniche di classificazione

L'analisi dei gruppi, o cluster analysis, ha lo scopo di produrre dei raggruppamenti di linee o di colonne della matrice dei dati d'interesse e di caratterizzarli statisticamente. Si tratta di oggetti o di individui descritti da un certo numero di variabili (quantitative o qualitative o di entrambi i tipi).

I raggruppamenti vengono effettuati mediante precise tecniche di classificazione che tengono conto delle *somiglianze* tra le unità statistiche, o equivalentemente delle loro *distanze*, in modo che

a) le unità riunite nello stesso gruppo siano molto simili o comunque vicine rispetto al complesso delle loro caratteristiche;
b) le unità appartenenti a gruppi diversi siano dissimili e lontane.

Le tecniche di raggruppamento operano solitamente su una *matrice di dissimilarità*, che contiene informazioni sul grado di dissomiglianza tra le diverse unità statistiche, e permettono di individuare una *partizione* o una *sequenza di partizioni* dell'insieme iniziale.

Le circostanze d'impiego delle metodologie di analisi dei gruppi sono sostanzialmente le stesse di quelle dei metodi d'analisi fattoriale. I dati di riferimento possono trovarsi raccolti

a) in una tabella contenente i valori numerici (solitamente continui) che le variabili assumono per ogni unità statistica;
b) in una tavola di contingenza;
c) in una tabella di valori presenza/assenza, a seconda che le unità statistiche siano o no in possesso di determinate caratteristiche opportunamente specificate dalle variabili;
d) in una matrice quadrata simmetrica di *indici di somiglianza* o di *distanze*.

Il ricorso alle tecniche di classificazione

a) può essere motivato da ipotesi di lavoro relative alle caratteristiche dai dati da analizzare: in questi casi sostanzialmente si suppone che per la natura stessa dei dati in esame siano sottesi ad essi dei raggruppamenti;
b) può derivare semplicemente dall'esigenza operativa di sintetizzare i dati in gruppi con elevato potenziale descrittivo e interpretativo.

Al di là delle specifiche motivazioni che possono portare alla scelta dell'analisi dei gruppi, il suo preciso scopo resta comunque quello di individuare e mettere in evidenza *classi*, o *gruppi*, di unità statistiche. La rappresentazione sintetica di tali gruppi avviene sotto forma di

a) *partizioni* degli insiemi studiati (linee o colonne della matrice iniziale dei dati);
b) *gerarchie di partizioni*;
c) *alberi di partizioni*: in questo caso il riferimento va alla *teoria dei grafi*.

A differenza delle analisi fattoriali, basate sulla soluzione di equazioni e quindi sugli usuali calcoli formali statistici, le tecniche di classificazione prevedono il ricorso ad *algoritmi ricorsivi* di diversi tipi:

a) gli algoritmi *partitivi* portano direttamente a delle partizioni dell'insieme dei dati: è il caso, per esempio, del metodo di aggregazione attorno ai *centri mobili* (prossimo Paragrafo);
b) gli algoritmi *ascendenti*, o *agglomerativi*, procedono alla costruzione di classi per agglomerazioni successive di oggetti (singole unità o gruppi) presi due a due e forniscono una gerarchia di partizioni;

c) gli algoritmi *discendenti*, o *divisivi*, procedono per divisioni successive dell'insieme complessivo e anch'essi possono fornire una gerarchia di partizione.

10.3 Aggregazione attorno ai centri mobili

Gli algoritmi di aggregazione non gerarchica pongono la loro attenzione su un numero *prestabilito* di gruppi, ottenuti secondo criteri di ottimizzazione che cercano di ottenere una buona compattezza all'interno dei gruppi e una buona separazione tra essi. Effettuata una partizione iniziale delle unità statistiche, si procede in modo iterativo attraverso *allocazioni* o *ri-allocazioni* delle unità stesse nei diversi gruppi.

Anche se fa riferimento a una base formale matematica piuttosto limitata e la sua validità è provata solamente dalle sue applicazioni pratiche, il metodo di *classificazione attorno ai centri mobili* è probabilmente la tecnica di raggruppamento attualmente più utilizzata ed è molto efficace sia in fase di analisi descrittiva che come tecnica di riduzione dei dati, generalmente in associazione ad analisi fattoriali.

Questo metodo, dovuto principalmente a Forgy, può essere considerato come un caso particolare delle tecniche conosciute sotto il nome di "*nubi dinamiche*", sviluppate formalmente da Diday.

I *centri di gravità*, o *centroidi*, dei gruppi vengono solitamente individuati attraverso i vettori delle medie aritmetiche – è il caso di questa Tesi – o delle mediane [1].

Dato un insieme I di n unità statistiche descritte da p variabili, definita una distanza d nello spazio p-dimensionale $\Re^p$ – in questa Tesi, la distanza secondo la metrica del χ^2 – e scelto il numero massimo q di gruppi da individuare, l'algoritmo prevede le seguenti fasi:

a) Passo 0 $\rightarrow$ Si determinano q centri provvisori dei gruppi, scelti con criteri soggettivi o con criteri casuali, per

[1] Così come avviene nell'ambito dell'ACM, anche nell'analisi dei gruppi viene immaginato un parallelo tra sistemi statistici e sistemi fisici.

esempio attraverso un'estrazione senza reimmissione. Questi centri iniziali ${}^0C_k$, con $k = 1,\ldots,q$, inducono una prima partizione ${}^0P$ dell'insieme I in q classi: l'unità i-esima appartiene al gruppo ${}^0I_k$ se è più vicina, in termini di distanza d, al centro ${}^0C_k$ che a tutti gli altri centri.

b) Passo 1 $\rightarrow$ Si determinano i nuovi centri di gravità ${}^1C_k$ dei gruppi ${}^0I_k$ ottenuti al passo 0. Questi nuovi centri inducono una nuova partizione ${}^1P$ dell'insieme I, sempre secondo la distanza d, nei gruppi ${}^1I_k$.

c) Passo $m \rightarrow$ A ogni ulteriore passo si ripetono le operazioni del passo 1, individuando i nuovi centri ${}^mC_k$ in base alla precedente partizione ${}^{m-1}P$ nei gruppi ${}^{m-1}I_k$ e ottenendo la nuova partizione ${}^mP$ nei gruppi ${}^mI_k$.

Al crescere del numero di passi il processo di classificazione tende a stabilizzarsi. Infatti tra il generico passo m e il passo m+1 la varianza entro i gruppi non può che decrescere o restare stazionaria; opportune *regole di assegnazione* delineate in sede di informatizzazione dell'algoritmo permettono di ottenere una descrescenza stretta, da cui segue immediatamente il processo di convergenza, dal momento che l'insieme di partenza I è finito.

L'algoritmo termina quando

a) due iterazioni successive portano alla stessa partizione, che viene quindi considerata definitiva;

b) una misura di variabilità opportunamente scelta (solitamente la varianza o l'inerzia tra i gruppi) termina di decrescere in modo statisticamente significativo;

c) si stabilisce un numero massimo d'iterazioni e viene raggiunta tale soglia.

La partizione finale dipende dalla collocazione iniziale dei centri (scelti o estratti).

10.4 Il metodo delle k medie

Il *metodo delle k medie* è dovuto a Mac Queen e fa parte delle tecniche note come "*nearest centroid sorting*" (classificazione del centroide più vicino): i centri di gruppo sono individuati da un insieme di "*semi*" (le cui caratteristiche variano a seconda del metodo adottato) e le unità statistiche vengono assegnate ai gruppi con il seme più vicino.

Mentre con il metodo dei centri mobili i nuovi centroidi vengono calcolati solamente dopo avere eseguito la ri-allocazione di *tutte* le unità nei gruppi, con il metodo di Mac Queen i centroidi vengono ricalcolati sulla base dei membri "correnti" dei gruppi, cioè a ogni ri-allocazione.

L'algoritmo prevede i seguenti passi:

1) Date *n* unità, si considerano le prime *k* come costitutive di *k* gruppi di un'unità ciascuno.
2) Si assegna ognuna delle rimanenti *n-k* unità al gruppo con centroide più vicino. Dopo ogni allocazione si ricalcola il centroide del gruppo a cui è stata aggiunta l'unità.
3) Quando tutte le unità sono state classificate si considerano i centroidi dei gruppi individuati come nuovi semi e si procede a una ri-allocazione delle unità sulla base dei semi più vicini.

Questo metodo si distingue per la sua "economicità" poiché utilizza le prime *k* unità come semi (evitando una scelta soggettiva o un'estrazione) e opera una sola ri-allocazione. Lo "sforzo computazionale" richiede solamente $k \cdot (2n - k)$ calcoli di distanze, $(k - 1) \cdot (2n - k)$ confronti di distanze e $(n - k)$ centroidi.

Anche in questo caso, però, la classificazione non è unica ma dipende dalla disposizione dei primi *k* centri di gravità.

10.5 La classificazione gerarchica e il metodo di Ward

Gli algoritmi gerarchici aggregativi assumono come situazione di partenza una configurazione in cui ciascuna unità costituisce un gruppo a sé stante.

Il passo successivo consiste nell'aggregare tra loro i 2 gruppi meno dissimili, ottenendo una ripartizione degli n elementi iniziali in n-1 gruppi di cui uno è composto da 2 unità.

Il processo viene iterato n-1 volte fondendo assieme di volta in volta i 2 gruppi meno dissimili, finché non si perviene alla collocazione di tutte le n unità in un solo gruppo (ritorno all'insieme iniziale).

Il prodotto finale dell'algoritmo è una serie completa di *partizioni concatenate*. La classificazione in uno specifico numero di gruppi è vincolata a quella relativa a un numero di gruppi superiore: una volta allocate nello stesso gruppo a un dato livello del processo, 2 unità non possono più essere separate successivamente. Ciò comporta l'impossibilità di perfezionare le assegnazioni già effettuate e costituisce un limite dei metodi aggregativi gerarchici.

Le tecniche gerarchiche di classificazione si differenziano per i diversi criteri che regolano la valutazione delle distanze o delle dissomiglianze.

Il *metodo di Ward* fa riferimento alla *devianza tra i gruppi*.

Dato un insieme I di n elementi in uno spazio geometrico-statistico q-dimensionale, sia P_{S+1} una partizione di I in S+1 sottoinsiemi $C_1,C_2,\ldots,C_{S+1}$. Per passare da una partizione P_{S+1} a una partizione P_S occorre fondere quei 2 gruppi C_a e C_b di P_{S+1} nel gruppo C_r di P_S facendo in modo che sia minimo l'incremento Δ

della devianza *entro* C_r [2], calcolata come somma dei quadrati delle distanze degli elementi del gruppo C_r dal proprio centro di gravità.

Si dimostra che se si fondono C_a di n_a elementi e C_b di n_b elementi nel nuovo gruppo C_r l'incremento Δ è proporzionale al quadrato della distanza tra i centroidi dei gruppi originari:

$$\Delta = \frac{n_a \cdot n_b}{n_a + n_b} \sum_{i=1}^{q} \left(\bar{x}_{i,C_a} - \bar{x}_{i,C_b} \right)^2 . \qquad (10.1)$$

La scelta delle coppie di gruppi da aggregare si basa quindi, a ogni passo dell'algoritmo, sulla minimizzazione della distanza tra i loro centroidi, che corrisponde alla minimizzazione della devianza *tra*.

[2] La fusione di 2 unità o di 2 gruppi comporta sempre un incremento della varianza e della devianza entro i gruppi. Ovviamente nella situazione di partenza con i gruppi costituiti da una sola unità statistica la varianza e la devianza entro i gruppi coincidono e sono nulle.

10.6 L'analisi dei gruppi con la classificazione mista

Il metodo d'aggregazione attorno ai centri mobili e il metodo delle k medie offrono vantaggi incontestabili perché consentono di ottenere partizioni su un insieme voluminoso di dati con un basso "costo operativo", ma nello stesso tempo presentano due inconvenienti:

a) producono partizioni che sono dipendenti dalla collocazione dei primi centri;
b) richiedono di fissare a priori il numero di gruppi.

Al contrario, la classificazione gerarchica dà luogo a una famiglia di algoritmi che possono essere definiti "deterministi", in quanto portano sempre allo stesso unico risultato se ri-applicati allo stesso insieme di dati, e fornisce informazioni per la scelta del numero di gruppi, che non va stabilito a priori. Le tecniche gerarchiche, però, presentano lo spiacevole inconveniente di essere poco adatte alla classificazione di vasti insiemi di dati.

Sembra allora vantaggioso ricorrere alla *classificazione mista*, combinando le diverse tecniche di classificazione illustrate nei Paragrafi precedenti nel tentativo di valorizzarne i pregi evitandone invece i difetti. Questa scelta metodologica è stata messa spontaneamente in opera da numerosi ricercatori, specialmente a partire dagli anni Ottanta.

Le più diffuse tecniche di analisi dei gruppi effettuata con metodi di classificazione mista prevedono sostanzialmente 4 fasi:

1ª fase: individuazione dei gruppi stabili

Viene effettuata una *partizione iniziale* in un numero piuttosto elevato di gruppi omogenei (qualche decina o anche – con migliaia di unità statistiche – qualche centinaio) attraverso il metodo dei centri mobili, delle nubi dinamiche o delle k medie.

Dal momento che queste tecniche individuano partizioni dipendenti dalla disposizione iniziale dei centri o delle nubi, viene attivata una procedura di *auto-convalida*: si incrociano diverse possibili partizioni iniziali e si ottengono come classi finali i "*gruppi stabili*", detti anche "*forme forti*", costituiti dagli individui che risultano essere stati *sempre raggruppati assieme* in tutte le partizioni effettuate.

2ª fase: costruzione dell'albero gerarchico

Si procede poi a una classificazione gerarchica, per esempio attraverso il metodo di Ward, dei gruppi stabili individuati: questi vengono aggregati due a due fino ad ottenere un unico gruppo (ritorno all'insieme iniziale) che costituirà la cima dell'*albero gerarchico*, detto anche *dendrogramma*.
L'analisi del dendrogramma può suggerire il numero finale di gruppi da considerare. Ogni taglio dell'albero individua una partizione avente tanti meno gruppi quanto più esso è vicino alla cima.

3ª fase: indici di livello, consolidamento e partizione ottimale

Si sceglie il livello o un ventaglio di possibili livelli in cui *tagliare* l'albero gerarchico.

Approssimandosi alla cima del dendrogramma, la distanza tra le classi da aggregare è via via maggiore. Si possono allora ordinare i nodi dell'albero in funzione di tale distanza e in riferimento ad aumenti significativi di essa. A tale scopo si analizza l'andamento di un opportuno *indice di livello*: tagliando l'albero in corrispondenza di un salto importante di questo indice si ottiene una partizione presumibilmente di buona qualità, dal momento che i gruppi individuati in base al taglio saranno sensibilmente distanti.

Non si ha però la garanzia che una partizione di questo tipo sia anche la migliore possibile. L'algoritmo di aggregazione, infatti, fornisce solamente un ottimo *locale*: la partizione ottenuta è la migliore *data* la partizione precedente. La classificazione può allora essere migliorata mediante una *procedura di consolidamento*, che opera una *ri-allocazione delle unità statistiche in modo da rendere minima l'inerzia entro le classi*: effettuato il taglio dell'albero è dunque conveniente procedere a un'ottimizzazione della classificazione applicando nuovamente il metodo dei centri mobili o delle nubi dinamiche o delle k medie e minimizzando la funzione-obiettivo "inerzia entro".

È inoltre consigliabile provare con diversi tagli dell'albero e calcolare per ognuno di essi il rapporto fra l'inerzia *tra* le classi e l'inerzia *totale*. Con l'applicazione della procedura di consolidamento si ottiene un incremento del valore di tale rapporto, in quanto viene minimizzata l'inerzia entro i gruppi e quindi aumenta l'inerzia tra i gruppi. La *partizione ottimale* sarà allora data dalla combinazione ritenuta migliore tra l'incremento del valore inerzia tra / inerzia totale e il numero di gruppi.

4ª fase: caratterizzazione dei gruppi

Scelta la partizione definitiva si arriva finalmente alla *caratterizzazione statistica dei gruppi ottenuti*, consistente nell'individuazione delle modalità e delle variabili più rilevanti ai fini di un'esauriente descrizione dei singoli gruppi.

La caratterizzazione viene operata in riferimento

a) a *statistiche di scarto* tra le frequenze delle modalità all'interno dei gruppi e le corrispondenti frequenze nell'insieme iniziale di unità statistiche;
b) ai *valori-test* delle modalità, illustrati nel prossimo Paragrafo.

10.7 L'analisi dei valori-test

Una modalità non risulta caratterizzare un dato gruppo se è presente nel gruppo stesso con una frequenza relativa che non si discosta in modo statisticamente significativo dalla frequenza relativa assunta nell'insieme complessivo delle unità.

In termini inferenziali, parlare di differenze statisticamente non significative, rispetto a una data modalità, tra la frequenza relativa nel gruppo e quella nella "popolazione" di riferimento equivale a ipotizzare che le unità statistiche coinvolte nel gruppo siano estratte casualmente da tale popolazione. Per ogni *k*-esimo gruppo, quindi, l'ipotesi nulla H_0 da controllare è l'ipotesi di estrazione casuale senza reintroduzione di n_k unità tra le n della popolazione di riferimento. Indicata con n_{jk} la numerosità delle unità che nel gruppo *k*-esimo possiedono la modalità *j*-esima, sotto questa ipotesi vale l'uguaglianza

$$\frac{n_{jk}}{n_k} = \frac{n_j}{n} \quad . \tag{10.2}$$

L'ipotesi alternativa H_1 specifica invece che la proporzione di unità che presentano la *j*-esima modalità è *più elevata* nel *k*-esimo gruppo rispetto al collettivo:

$$H_1: \frac{n_{jk}}{n_k} > \frac{n_j}{n} \quad . \tag{10.3}$$

Più ci si allontana da H_0 e diventa verosimile H_1, più la modalità è da considerarsi importante nella caratterizzazione del gruppo.

La numerosità n_{jk} viene a corrispondere a una variabile aleatoria N che sotto l'ipotesi H_0 segue una legge di probabilità ipergeometrica dai parametri noti. Essendo la distribuzione ipergeometrica convergente alla distribuzione binomiale e di

conseguenza a quella normale [Paruolo 1992], i *valori-test* sono ricavati dalla normale standardizzata e individuano le probabilità critiche

$$p_{jk} = Prob\ \{N >= n_{jk} \mid H_0\}\ . \qquad (10.4)$$

All'aumentare del valore-test diminuisce la probabilità critica e diventa sempre più inverosimile l'ipotesi nulla.

Analogamente al caso dei valori-test per le modalità illustrative nell'ACM (Capitolo 8), nel prossimo Capitolo saranno considerate significative le modalità che portano a valori-test maggiori di +2 o minori di -2, corrispondenti a una probabilità critica inferiore a circa il 5% [[3]].

[3] Impostazione di default di SPAD.