1. Ricerca e individuazione di gruppi

1.1 Classificazione, tipologie, tassonomie

Nel parlare di analisi dei gruppi è possibile fare una distinzione tra due situazioni:

a. ricerca di gruppi omogenei indipendentemente dalla loro reale esistenza (analisi tipologica): in questo caso non si cerca una conformità a modelli teorici;
b. verifica di gruppi già esistenti attraverso le variabili considerate e gli algoritmi adottati.

Nel primo caso l'analisi adotta metodi detti di *cluster analysis* mentre nel secondo si applicano metodi confermativi quali l'analisi discriminante (trattata in un altro capitolo). Per poter comprendere i diversi approcci all'analisi dei gruppi può essere utile definire alcuni concetti fondamentali.

Classificazione

In generale la classificazione rappresenta un processo centrale non solo in molti aspetti della vita quotidiana, ma anche in tutti gli ambiti scientifici che richiedono concettualizzazione, ragionamento e linguaggio avanzati. Rappresenta un importante approccio concettuale nell'ambito della matematica, della statistica e dell'analisi dei dati in generale. Si può parlare di classificazione quando le classi identificate sono *esaustive* e *mutuamente esclusive*. Le classi identificate possono essere

– *monotetiche*, quando contengono elementi identici relativamente alle variabili identificate;
– *politetiche*, quando non contengono elementi identici relativamente alle variabili identificate, ma piuttosto raggruppano elementi somiglianti.

Quando la classificazione non è osservabile *in natura*, è possibile procedere in modo empirico utilizzando un sistema artificiale. In questi casi il termine classificazione può riferirsi sia al processo che al risultato. In termini di procedimento, la classificazione è definita semplicemente come l'ordinamento/raggruppamento di entità in gruppi o classi sulla base della loro somiglianza. Al termine del procedimento, le entità appartenenti allo stesso gruppo dovrebbero essere il più possibile omogenee tra di loro e il più possibile diverse da quelle appartenenti agli altri gruppi individuati. Statisticamente ciò può essere definito facendo riferimento alla varianza: i gruppi dovrebbero essere individuati in modo da minimizzare la varianza *all'interno* del gruppo (*within*) e massimizzare la varianza *tra* gruppi (*between*). La classificazione avviene in base ad una o più caratteristiche/variabili. Quando se ne utilizzano diverse, si assume che tra le variabili esista una relazione che giustifichi il procedimento. La scelta delle variabili rappresenta un momento chiave del procedimento di classificazione. Infatti, anche se la classificazione viene effettuata seguendo alla perfezione le regole poste, la selezione di variabili irrilevanti produrrà comunque

Filomena Maggino, *L'analisi dei dati nell'indagine statistica.*
Volume 2: l'esplorazione dei dati e la validazione dei risultati,
ISBN: 88-8453-351-1 (print) ISBN: 88-8453-350-3 (online),

classificazioni insignificanti. Nel caso più fortunato, le variabili selezionate possono produrre classificazioni in modo molto semplice. D'altra parte la classificazione empirica non sempre è semplice e presenta spesso grossi problemi. La difficoltà di raggruppare attraverso elementi di somiglianza cresce in modo esponenziale con il numero degli oggetti da classificare, con il numero di variabili che devono essere prese in considerazione e con il numero di categorie/modalità che definiscono la misurazione di ciascuna variabile.

Confrontare due oggetti per volta in riferimento a molte variabili può diventare un compito particolarmente difficile nel caso in cui il numero degli oggetti sia particolarmente alto.

L'introduzione del computer nell'analisi statistica ha consentito di alleggerire non poco tale compito con la possibilità di introdurre particolari algoritmi e formule adatte nei casi in cui si disponga di campioni numerosi e di un elevato numero di variabili.

Le procedure di classificazione racchiudono tre livelli di analisi:

a. livello concettuale (classificazione di concetti),
b. livello empirico (classificazione di entità empiriche),
c. livello operazionale o di indicatori (classificazione che combina i due livelli precedenti).

Il procedimento di classificazione può essere:

– *sincronico* o *cross-sectional* o *non evolutive* ovvero si verifica in un singolo punto di tempo; in biologia questi casi si definiscono relazioni *fenetiche*;
– *diacronico*, quando è basato su misure di cambiamento o su misure di somiglianza evolutiva; in biologia questi casi si definiscono relazioni *filetiche* (*phyletic*) e mostrano il corso di una evoluzione.

Tipologia

Una tipologia può essere vista come classificazione concettuale. Le combinazioni che definiscono una tipologia rappresentano concetti-tipo piuttosto che casi empirici. Le tipologie sono caratterizzate da etichette o nomi. Si ponga il caso di avere due caratteristiche sulla base delle quali costruire delle tipologie (autostima, motivazione) per ciascuna delle quali sono state identificate due classi (con/senza autostima; motivato/non motivato). Combinando tra loro le categorie è possibile definire quattro tipologie. Il numero delle tipologie aumenta all'aumentare delle caratteristiche e delle categorie che le definiscono[1]. È anche possibile che, nel procedere alla combinazione delle caratteristiche e delle categorie, alcune tipologie risultino concettualmente inaccettabili.

In genere, ma non sempre, le tipologie contengono solo classi monotetiche; in altre parole, un tipo rappresenta una classe monotetica.

[1] Basti pensare che cinque caratteristiche con due categorie producono 32 tipologie, mentre 12 ne producono 2^{32} (4096). Nel caso di caratteristiche con un numero di categorie maggiore, come spesso succede, il numero delle tipologie si espande ancora più rapidamente.

Tassonomia

Come la *classificazione*, anche il termine *tassonomia* può essere riferito sia al processo che al risultato finale. La tassonomia, vista in termini di processo, è definita come "lo studio teorico che comprende le basi, i principi, le procedure e le regole della classificazione", cui altri aggiungono lo studio teorico dell'identificazione[2].

La tassonomia, vista in termini di risultato finale, è simile alla tipologia; è per questo che molti utilizzano i due termini in modo intercambiabile. La principale differenza sta nel fatto che la *tipologia*, più utilizzata nelle scienze sociali, è concettuale mentre la *tassonomia* è empirica e utilizzata soprattutto nelle scienze biologiche, nell'ambito delle quali le tassonomie sono spesso (ma non sempre) create in termini gerarchici (es.: famiglia → genere → specie) ed evolutivi.

Una cella di una tassonomia è detta *taxon*; più celle sono dette *taxa*.

Nel procedere ad una classificazione è importante tenere presente che essa presenta importanti vantaggi ma anche degli svantaggi; vediamone alcuni.

Vantaggi

Vi sono diversi vantaggi che rendono la classificazione non solo utile ma anche, in alcuni casi, necessaria:

1. *Descrizione e confronto*. Una buona classificazione consente anche di descrivere e confrontare gruppi in modo veloce e facile.
2. *Riduzione di complessità* e raggiungimento della *parsimonia*: nei casi di popolazioni (di individui, oggetti, concetti, ecc.) molto numerose è difficile giungere ad una corretta ed esaustiva descrizione; la classificazione consente di semplificare la complessità della realtà in modo sufficiente da consentire di analizzarla sulla base delle caratteristiche considerate.
3. *Analisi di casi sulla base di somiglianze e differenze*: la classificazione, identificando e raggruppando casi simili, consente di fare particolari analisi; un tipico esempio è rappresentato dall'analisi diagnostica in medicina: l'identificazione di casi simili consente di identificare gruppi di persone che soffrono degli stessi sintomi; ciò può condurre al riconoscimento della malattia sottostante.
4. *Definizione di criteri di misurazione*: una buona definizione di classificazione consente di determinare semplici strumenti di misurazione.

Svantaggi

Anche se l'importanza della classificazione è largamente riconosciuta in molte discipline scientifiche, le scienze sociali si sono mostrate sempre critiche, soprattutto per i seguenti motivi.

1. *La classificazione non è esplicativa*: pur essendo un utile strumento descrittivo, la classificazione non consente alcuna spiegazione e/o previsione; d'altra parte qualsiasi spiegazione/previsione è difficile se non è basata o sostenuta da un adeguato sistema descrittivo quale la classificazione.

[2] L'*identificazione* rappresenta il procedimento di ricerca dei casi empirici per ciascuna cella.

2. *Reificazione*: la classificazione creata teoricamente (definizione di tipi ideali e costruiti) può essere confusa con una entità empirica e corre il pericolo di venire trattata come reale.
3. *Staticità*: la classificazione è soprattutto sincronica e statica, ciò la rende insufficiente nella descrizione di sistemi dinamici e diacronici.
4. *Difficoltà nella identificazione di casi e di variabili*.
5. *Ingestibilità*: mentre la classificazione semplice può risultare di poco valore applicativo, la classificazione complessa può risultare di difficile applicazione; tale critica viene superata dalla possibilità di utilizzare strumenti computerizzati; d'altra parte, uno degli obiettivi della classificazione è proprio quello di ridurre la complessità ed assicurare la gestibilità.

1.2 La *cluster analysis*

Uno dei più interessanti approcci all'analisi dei dati è senz'altro quello orientato alla ricerca e all'individuazione di gruppi. Una metodologia di analisi particolarmente flessibile finalizzata alla soluzione di problemi di classificazione e che ha l'obiettivo di organizzare gli elementi (soggetti, cose, eventi, ecc.) in strutture significative (dette *cluster*[3])è quella che va sotto il nome di *cluster analysis*. Secondo questo approccio i gruppi sono definiti in modo tale che

- gli elementi appartenenti allo stesso gruppo risultino molto simili tra loro, mentre quelli appartenenti a gruppi distinti risultino tra loro molto diversi, ovvero
- il livello di associazione sia stretto tra i componenti dello stesso *cluster* e molto debole tra gli elementi appartenenti a *cluster* diversi.

In questo senso la *cluster analysis* rappresenta uno strumento di analisi *esplorativa*, in quanto può mettere in evidenza associazioni e strutture nei dati non altrimenti rilevabili e che possono risultare utili una volta individuate.

Spesso il procedimento viene definito "automatico" perché l'individuazione dei gruppi è raggiunta tramite algoritmi formalizzati. La definizione di gruppi omogenei è utile in tutti quei casi in cui vi sia la necessità di:

– ridurre la complessità dei dati rispetto alle unità, identificando e descrivendo forti connessioni tra i casi (tipologie);
– riunire i dati in maniera significativa e per mezzo di metodi quantitativi;
– scoprire i legami esistenti tra casi;
– costruire sistemi di classificazione automatica che consentono di immagazzinare informazioni, documenti, ecc.; nelle scienze biologiche ciò viene definito *tassonomia*;
– esplorare i dati in una forma grafica che sia
  - semplice, che metta in evidenza le informazioni dei dati,
  - sintetica, in quanto rappresenta i risultati in poche dimensioni;

[3] In inglese *cluster* vuol dire *grappolo*, *gruppo*, *sciame*.

– attribuire ai casi che presentano dati mancanti, valori noti attraverso la conoscenza del gruppo cui tali casi appartengono per omogeneità;
– stratificare popolazioni da sottoporre a campionamento;
– studiare gli effetti di diversi trattamenti sperimentali;
– formulare e verificare ipotesi di classificazione dei casi al fine di identificare l'eventuale presenza di modelli.

Tale metodologia di analisi trova applicazione in diversi settori scientifici:
– in archeologia (per la classificazione di strumenti di epoche diverse o l'assemblaggio di reperti per una loro datazione o per spiegare la loro origine culturale),
– in demografia (per dedurre conseguenze sociali e genetiche di movimenti migratori),
– in economia (per classificare regioni e identificare aree omogenee sulla base di particolari indici),
– in psicometria (per stabilire la validità di una certa classificazione di soggetti o di variabili, verificare ipotesi già formulate per il raggruppamento di *item* in aree o per la definizione di tipologie di soggetti, per analizzare e strutturare i giudizi dei soggetti),
– in psicologia dell'educazione (per classificare studenti al fine di definire modelli e moduli didattici differenziati),
– in geografia (per individuare raggruppamenti territoriali sulla base di attività agricole o per misurare cambiamenti nella produttività),
– nelle scienze giuridiche (per classificare e derivare classificazioni operative delle leggi all'interno, per esempio, della comunità europea),
– in biblioteconomia (per classificare le pubblicazioni attraverso parole-chiave e citazioni, per produrre raggruppamenti al fine dell'identificazione, l'estrazione, l'aggiornamento di titoli),
– in linguistica (per tracciare l'evoluzione e lo sviluppo di una tassonomia semantica),
– nella scienza politica (per raggruppare i risultati delle votazioni ed eventualmente cercare di prevedere possibili risultati futuri),
– nelle scienze sociali (per classificare particolari tendenze, atteggiamenti, comportamenti e ruoli sociali per sviluppare modelli tipologici di classi e identità sociali),
– in antropologia (per la classificazione di dati antropometrici per rilevare differenze genetiche tra razze umane al fine di ricostruire la loro evoluzione),
– in biochimica (per raggruppare la composizione degli amino-acidi delle proteine e dei geni, per spiegare la sequenza evolutiva delle mutazioni che hanno originato nuove specie),
– in botanica (per classificare campioni vegetali, per descrivere l'ecologia di comunità naturali e indicare aree per lo sviluppo o la conservazione agricola, per studiare metodi di prevenzione dell'erosione del suolo),
– in psicologia clinica (per identificare e classificare modelli di comportamento sociale estremi, per identificare sindromi associate con particolari patologie),
– in citologia (per classificare campioni ematici, per definire gruppi sanguigni e tipologie di plasma e sviluppare metodi per verificare la presenza di cellule anormali),
– in medicina (per fare diagnosi di quadri clinici, previsioni di morbilità di individui e popolazioni),

– nelle scienze naturali (per affrontare problemi di tassonomia),
– in psichiatria (per classificare sintomi in modo da identificare sindromi di disordini psichiatrici e sviluppare trattamenti),

e in biometria, microbiologia, ingegneria, ecc.

Lo sviluppo della *cluster analysis* non è avvenuto nell'ambito di una singola disciplina; è per questo motivo che spesso soluzioni e metodi simili vengono indicati con nomi diversi: ciò può aver prodotto una certa sovrastima dei metodi effettivamente disponibili.

Definizione di *cluster*

La definizione di *cluster* è di carattere pragmatico in quanto consente di stabilire, come si vedrò, se più elementi formano un gruppo[4]. Ad un livello superficiale, l'idea alla base di tutti i metodi di *cluster analysis* appare molto semplice; l'obiettivo è quello di cercare di raggruppare oggetti in gruppi simili, detti *cluster*, in modo tale che i componenti di un gruppo siano simili tra loro e meno simili con i componenti degli altri gruppi.

Una possibile definizione è quella che implica l'applicazione di alcuni concetti statistici.

Identificato un gruppo di casi è possibile individuare e calcolare il centro del *cluster* ("centroide") sulla base dei valori registrati dai casi per le variabili considerate; se la variabile è una, il centro del gruppo equivale al punteggio medio degli stessi casi per quella variabile.

Identificati due distinti *cluster* relativamente ad una variabile, le medie e le varianze calcolate nei due gruppi dovrebbero risultare piuttosto diverse. In particolare la varianza di ciascuno dei gruppi dovrebbe essere minore della *varianza totale* ottenuta combinando i due gruppi. Si può anche dire che la varianza "all'interno" (*within*) dei *cluster* dovrebbe essere minore della varianza tra (*between*) *cluster*.

Occorre comunque dire che in letteratura il concetto di *cluster* non è definito in modo univoco; in genere sono le strategie e i singoli algoritmi per le applicano che consentono di dare una definizione pratica e implicita.

1.2.1 Il procedimento di analisi

Per poter realizzare un'analisi di raggruppamento il ricercatore deve stabilire

1. su quale tipo di elementi della matrice dei dati intende procedere al raggruppamento (in genere, sui casi),
2. rispetto a quali variabili ricercare il raggruppamento,
3. con quale tipo di logica procedere al raggruppamento (strategie di *clustering*),
4. a partire da quale matrice di prossimità (distanze o somiglianze) tra casi procedere al raggruppamento,
5. quale tecnica adottare per eseguire la strategia; alla tecnica è legata anche la scelta della misura di prossimità tra gruppi.

[4] In molte discipline si ricorre a definizioni pragmatiche: in fisica per esempio non si definisce la temperatura, ma come la si misura, in statistica non si definisce la probabilità, ma gli assiomi che deve soddisfare.

1.2.1.1 Identificazione delle variabili per la classificazione

La scelta delle variabili da utilizzare è naturalmente legata agli obiettivi dell'analisi e richiede la definizione di un modello logico. Tale scelta dovrebbe ricadere su quelle variabili che descrivono il fenomeno relativamente al quale deve essere svolta l'analisi. Quindi all'interno di una determinata matrice di dati vengono selezionate quelle variabili che si ritengono significative per l'identificazione dei *cluster*. Nella scelta occorre tenere presente che se si utilizzano variabili con basso potere discriminante, l'analisi di raggruppamento può non produrre differenze significative tra le unità; al contrario, l'utilizzo di variabili con alto potere discriminante può rendere inutile l'inclusione delle altre variabili che logicamente sono molto legate al fenomeno.

L'importanza relativa di ciascuna variabile nella formazione dei gruppi è collegata alla varianza delle diverse variabili, da qui l'utilità e l'importanza di standardizzare, tranne che nel caso di variabili dicotomiche, le variabili.

1.2.1.2 Strategie per l'individuazione dei gruppi

Per poter identificare dei *cluster* in un gruppo di dati è necessario stabilire una strategia. Una prima distinzione che può essere fatta è quella che distingue tra (figura I. 1.1)[5]:

- strategia che ammette *cluster sovrapposti*: in questo caso si ammette che uno stesso elemento possa comparire in uno o più *cluster* (strategia di *clumping*);
- strategia che ammette *cluster esclusivi*: in questo caso ogni elemento può comparire in un solo *cluster*.

Nell'ambito della strategia esclusiva è possibile distinguere tre diversi approcci:

- Soluzione gerarchica: i gruppi sono individuati in fasi successive secondo livelli ordinati; alla fine di tale procedimento si ottiene una successione di raggruppamenti sempre meno differenziati ottenendo così una struttura di tipo piramidale; in pratica è possibile distinguere due diverse logiche gerarchiche:

[5] Secondo Bailey (1994) è possibile identificare ben 15 criteri di classificazione dei metodi di raggruppamento:

1. Strategia divisiva e strategia agglomerativa.
2. Metodi monotetici e metodi politetici.
3. Gruppi naturali e gruppi artificiali.
4. Metodi con numero di *cluster* predeterminati o no.
5. Tecniche *single level* e gerarchiche.
6. Tecniche per *cluster* sovrapposti ed esclusivi.
7. Outlier permessi o no.
8. Forma del legame.
9. Livello di somiglianza oggettiva e soggettiva.
10. Metodi combinatori e non combinatori.
11. Metodi compatibili e non compatibili.
12. Metodi iterativi e non iterativi.
13. Metodi sequenziali e metodi simultanei.
14. Metodi locali e globali.
15. Raggruppamenti pesati e non pesati.

Occorre però dire che, in pratica, solo alcuni di tali criteri vengono considerati; in particolare, nell'ambito dell'analisi dei dati di indagini sociali si considerano principalmente i criteri 1, 4, 5, 6, 8 e 9.
La maggior parte delle applicazioni utilizza un tipo di approccio che Sneath e Sokal (1973) hanno definito SAHN ovvero *Sequential, Agglomerative, Hierarchical, Nonoverlapping clustering methods* (metodi di raggruppamento sequenziali, agglomerativi, gerarchici, non sovrapposti).

- a partire da *n* gruppi composti da una sola unità si giunge, attraverso successive aggregazioni delle unità o dei gruppi più simili tra loro, alla formazione di un unico gruppo formato da *n* unità (*aggregazione gerarchica ascendente*); l'applicazione di tale logica richiede la definizione di *tecniche agglomerative*;
- a partire da un unico gruppo composto da *n* elementi, si arriva alla individuazione di *n* gruppi (*aggregazione gerarchica discendente*); l'applicazione di tale logica richiede la definizione di *tecniche divisive*.

La rappresentazione dell'intero procedimento giunge a descrivere un "albero di aggregazione"; tale albero consente al ricercatore di individuare a quale livello del procedimento si è raggiunta l'aggregazione più significativa. Questa strategia non richiede la definizione preventiva del numero di gruppi da ottenere.

– Soluzione non gerarchica: i gruppi sono individuati aggregando gli elementi in un numero prestabilito di *cluster* ottimizzando una funzione, detta *obiettivo* o *criterio*, che prende in considerazione le distanze tra i gruppi e/o tra le unità all'interno dei gruppi; si procede iterativamente prima identificando *r* gruppi (con $r < n$ e determinato a priori) in maniera casuale; successivamente si effettuano spostamenti delle unità tra i gruppi al fine di ottimizzare il criterio, ad esempio rendere massima la distanza tra i gruppi o minima quella tra le unità all'interno degli stessi. Naturalmente questa strategia richiede la definizione preventiva del numero di gruppi da ottenere.
– Soluzione additiva: note con il termine *Additive Trees*, che utilizzano tecniche grafiche di rappresentazione in cui le distanze dei "rami" riflettono le somiglianze tra gli oggetti.

Fig. I. 1.1 Strategie di clustering

Esistono anche strategie che applicano sia la logica gerarchica sia quella non gerarchica; altre strategie possono essere finalizzate all'individuazione di gruppi attraverso approcci specifici come quella finalizzata all'identificazione di particolari concentrazioni (*densità*); esistono inoltre tecniche miste.

1.2.1.3 Costruzione della matrice di prossimità tra unità

L'analisi di raggruppamento prende avvio dalla matrice di prossimità; tale matrice, come è noto, è simmetrica ed è calcolata a partire dalla matrice dei dati. La scelta della misura di prossimità è importante in quanto diverse misure possono condurre a risultati differenti. A tale proposito è opportuno ricordare che la misura della prossimità può essere definita in termini di distanza o di somiglianza.

Pur essendo definite e calcolate in modi diversi, le somiglianze e le distanze svolgono lo stesso compito. Nella maggior parte dei *package* statistici esiste la possibilità di calcolare tali misure; il ricercatore può comunque anche creare e sottoporre indici definiti in modo autonomo.

Nella scelta tra i diversi indici occorre tenere presente che

– le misure basate sulla correlazione non sono influenzate dalle differenze nelle scale di misura delle variabili e nelle dimensioni di tali valori nei casi (per esempio le somiglianze tra nazioni rispetto a determinate statistiche non sono influenzate dal fatto che alcuni stati hanno valori medi più grandi di altri);
– le altre misure, in particolare quelle basate su modelli euclidei o *city-block*, sono influenzate in maniera significativa dalle differenze nelle scale (per esempio due nazioni risulteranno essere diverse perché presentano valori molto diversi in senso assoluto pur seguendo modelli comuni).

È buona norma, prima di calcolare la matrice di prossimità, procedere alla standardizzazione:

- delle variabili, quando l'obiettivo è quello di raggruppare casi (righe della matrice dei dati) e le variabili utilizzate sono diverse tra loro rispetto alle grandezze utilizzate (scale di misurazione);
- dei profili, quando l'obiettivo è quello di raggruppare variabili (colonne della matrice) e i casi utilizzati sono diversi tra loro nei valori dei profili.

1.2.2 Le tecniche

Per poter realizzare una strategia è necessario disporre di una tecnica. Per ciascuna strategia è possibile identificare molte tecniche che spesso possono produrre soluzioni diverse. Il problema del ricercatore è quello di scoprire quale tra le diverse soluzioni è la più indicativa del raggruppamento "naturale" dei dati. A tale proposito occorre dire che tutte le tecniche producono sempre e comunque *cluster*, anche quando i dati non presentano alcun raggruppamento naturale. Ne consegue che il successo delle applicazioni di *cluster analysis* dipende completamente dal fatto di sapere se il modello di raggruppamento imposto corrisponde ad una struttura reale o meno.

Come si vedrà in seguito, sono molte le tecniche che consentono di applicare le diverse strategie di *cluster analysis* e non sempre si dispone di criteri oggettivi per la scelta tra i diversi approcci. Un criterio può essere quello che richiede la definizione e la verifica di caratteristiche ritenute desiderabili per la soluzione che si sta cercando.

Ma il criterio più importante di scelta è quello che fa dipendere la selezione dagli scopi dell'analisi, infatti ogni tecnica individua una partizione secondo diverse espressioni del concetto di "omogeneità" all'interno di ciascun *cluster*.

Inoltre per orientarsi nella scelta è importante riuscire a valutare i seguenti elementi:

- *oggettività*, ovvero possibilità di ripetere l'analisi in modo indipendente su un insieme di dati giungendo agli stessi risultati;
- *stabilità*, ovvero possibilità di applicare la classificazione su campioni equivalenti ottenendo risultati confrontabili; in questo senso si può orientare la scelta verso quelle tecniche che si presentano meno sensibili rispetto a piccole variazioni nei dati analizzati;
- *chiarezza* e *comunicativa* del risultato;
- *semplicità dell'algoritmo*, velocità di esecuzione.

Le diverse tecniche si distinguono principalmente per la strategia che realizzano. È possibile individuare anche altri criteri che consentono di distinguere le diverse tecniche; essi fanno riferimento a:

- Metodo di calcolo: i metodi di calcolo possono essere iterativi o non iterativi. Con i metodi iterativi un *cluster* viene continuamente migliorato in passaggi di calcolo successivi.
- Sequenza di individuazione dei *cluster*: in questo senso i metodi possono essere distinti in sequenziali e non sequenziali (o simultanei). Nei primi l'individuazione dei *cluster* procede in sequenza di passaggi anziché in un'unica operazione. Nei secondi l'individuazione dei *cluster* avviene in un unico passaggio. Quasi tutte le tecniche agglomerative sono sequenziali.
- Misure di prossimità utilizzate lungo il procedimento: in questo senso si può distinguere tra metodi compatibili e metodi non compatibili. I primi sono quelli per i quali le misure di prossimità calcolate lungo l'analisi sono sempre le stesse; i metodi non compatibili sono invece quelli in cui alcune proprietà delle misure originali vengono perdute lungo il corso dell'analisi. È evidente come in questo secondo caso sorgono seri problemi di interpretazione. La maggior parte dei metodi sono compatibili.
- Grado di affidabilità delle misure di prossimità: nei metodi gerarchici, le soluzioni di raggruppamento non sono uniformemente buone a tutti i livelli della gerarchia. La stima delle somiglianze tra oggetti può per esempio essere affidabile all'interno di un *cluster* ma può essere sempre meno affidabile al crescere del numero dei *cluster* considerati. Un tale tipo di metodo è detto *locale*.
- Peso da attribuire alle variabili utilizzate nelle procedure di *clustering*: esistono molte forme di pesi da utilizzare in questo tipo di analisi come per esempio considerare alcune variabili più importanti di altre. Secondo alcuni ricercatori qualsiasi forma di peso deve essere considerata arbitraria. Anche se tutte le procedure sono in pratica pesate, qualsiasi approccio che mira a dare pesi diversi deve essere presa in considerazione con molta cautela sia per motivi teorici sia pratici.

  Nella figura I. 1.2 sono sintetizzate le principali tecniche di *clustering*.

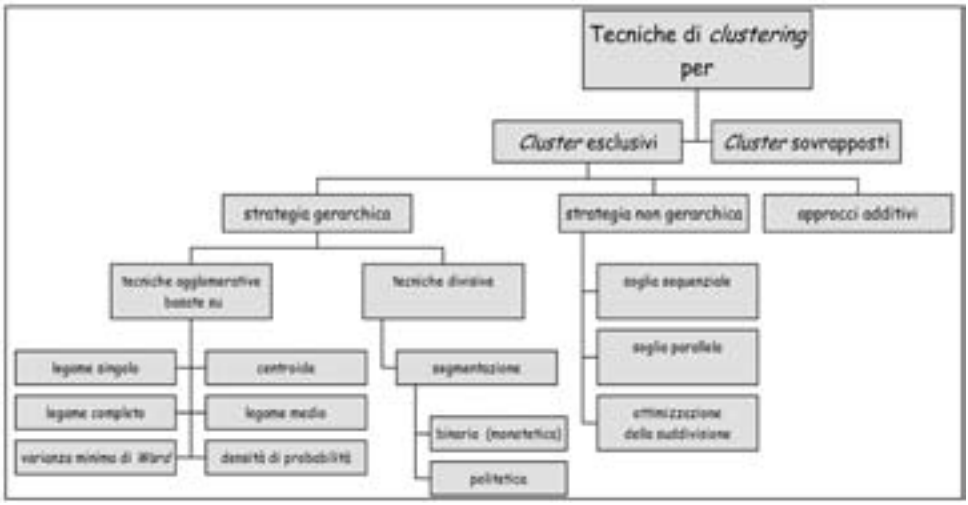

Fig. I. 1.2 Le tecniche di clustering

Tecniche per *cluster* sovrapposti

Anche se la individuazione di soluzione con classi sovrapposte difficilmente trova una soluzione statistica, sono state proposte alcune tecniche che ammettono la possibilità che, per un dato numero di gruppi, un'entità appartenga contemporaneamente a più di un raggruppamento separato. A questo gruppo appartengono le seguenti tecniche:

1. *insiemi sfocati (fuzzy set) unimodali*, utilizzato soprattutto in studi di tipo linguistico;
2. *miscugli di distribuzioni univariate o multivariate*; secondo tale tecnica ad ogni caso si associa la probabilità di appartenenza ai gruppi;
3. *analisi fattoriale Q*: metodo, utilizzato soprattutto in psicologia, che non è altro che una analisi fattoriale (v. capitoli successivi) applicata sulla matrice di distanza/somiglianza tra i casi; conseguentemente i *factor loading* non vengono associati alle variabili ma ai casi. La classificazione avviene assegnando ogni caso ad un gruppo sulla base del livello di saturazione dei fattori estratti. Per decidere a quale gruppo assegnare le unità è indispensabile una rappresentazione grafica delle entità sugli assi identificati dai fattori.

1.2.2.1 Tecniche di analisi per la strategia gerarchica

Tutte le tecniche che procedono secondo la strategia gerarchica sono iterative; esse possono essere distinte in tecniche divisive e tecniche agglomerative.

Tecniche per l'analisi gerarchica divisiva

Le tecniche divisive, dette anche *top-down*, procedono considerando le n unità come un unico insieme; tale insieme viene progressivamente suddiviso in $n-1$ passaggi giungendo, al termine del procedimento, alla situazione in cui ogni unità definisce un

gruppo. La suddivisione viene effettuata seguendo un certo criterio definito che cerca di ottimizzare la scissione.

La scissione può avvenire sulla base di:

– un attributo dicotomico alla volta (*segmentazione binaria* o *monotetica*)[6],

– tutto l'insieme degli attributi (*suddivisione politetica*).

Gli algoritmi che in genere vengono adottati, detti *scissori*, presentano particolari problemi applicativi in quanto, pur soddisfacendo le più rigorose proprietà statistico-matematiche, possono essere applicati ad un numero limitato di unità.

[6] La segmentazione binaria è un metodo *scissorio gerarchico* di partizione che mira a suddividere le unità osservate in gruppi il più possibile differenti tra loro, disponendo di una variabile quantitativa (y) e di un insieme di variabili (x), dette *predittori* e/o *esplicative*. A ciò si giunge mediante un procedimento iterativo costituito da una successione di progressive divisioni (*segmentazioni* basate sulle variabili esplicative) dicotomiche (*binarie*) di tipo gerarchico di uno dei gruppi di unità precedentemente formati in modo da minimizzare la varianza residua. Gli obiettivi di tale metodo possono essere così riassunti:

– classificare le unità in gruppi non predefiniti,

– individuare le variabili maggiormente esplicative e discriminanti della variabilità di y.

La procedura iterativa procede secondo vari stadi:

a. individuazione, per ogni unità, della *variabile spiegata* (y);
b. individuazione delle *variabili esplicative* (x_i);
c. determinazione per ogni variabile esplicativa di tutte le bipartizioni possibili;
d. per ciascuna bipartizionedi ciascuna variabile esplicativa, osservazione delle devianze nei gruppi (*within*) e quelle tra i gruppi (*between*) della variabile y;
e. analisi della bipartizione che produce la massima devianza *between*;
f. ripetizione del procedimento per ciascuna delle altre variabili esplicative.

Il procedimento dipende dall'ordine in cui si considerano le variabili. Con m variabili esplicative, i segmenti finali sono 2^m; a tale proposito occorre tener presente che alcuni di essi potrebbero contenere un numero molto piccolo di unità da non giustificarne la considerazione.

Un approccio analitico all'analisi di segmentazione binaria è l'*Automatic Interaction Detection* (*AID*). Le tecniche alla base dell'*AID* consentono di suddividere l'insieme delle unità (*gruppo genitore*) considerando tutte le possibili divisioni binarie sulla base di una sola variabile per volta. Viene scelta la partizione che minimizza la devianza nei gruppi (*within*) e massimizza quella tra gruppi (*between*) della variabile dipendente. Il procedimento si ripete su ognuno dei due gruppi *figli* ottenuti. È possibile bloccare l'analisi a diversi livelli prima che siano state esaminate tutte le variabili quando si raggiunge la dimensione minima dei gruppi, la minima capacità esplicativa della migliore suddivisione ad ogni passo, la minima devianza totale del gruppo genitore o il massimo numero di passi del procedimento.

Tenendo conto che una variabile x_i con a_i modalità produce $2^{a_i-1}-1$ possibili partizioni e che il numero totale delle suddivisioni da esaminare per tutte le variabili è $\sum 2^{a_i-1}-1$, se non si introduce un'ipotesi sulle aggregazioni possibili (ovvero sul numero di variabili e sul numero di categorie per ogni variabile), il procedimento di analisi può risultare molto elaborato e sterile.

Tale metodo presenta una diversa versione, detta *CHAID* e applicabile nel caso di variabili categoriche; tale versione è basata su una procedura detta *sequential merge-and-split* e sull'utilizzo del *chi-quadro*: dopo aver costruito una tabella incrociata tra le m categorie della variabile indipendente e le k categorie della variabile dipendente, si procede all'identificazione e all'aggregazione (*merge*) delle due categorie della variabile indipendente le cui sottotavole *2xk* risultano per il *chi-quadro* significativamente diverse; se il *chi-quadro* risulta non significativo rispetto al valore critico definito, si ripete il passaggio precedente per la variabile indipendente selezionata fino a quando non si presenta alcun risultato non significativo per il *chi-quadro* per una sottotavola; successivamente si identifica la variabile indipendente che presenta il valore *chi-quadro* maggiore e si suddivide (*split*) il gruppo in $m \leq l$ sottogruppi, dove l rappresenta il numero di categorie che risultano dal processo di *merging* effettuato su tale variabile; questo procedimento prosegue fino a quando non si osserva alcun risultato significativo di *chi-quadro*.

La maggior parte dei criteri di valutazione dei *cluster* ottenuti si basa sulla logica dell'analisi della varianza multivariata (MANOVA).

Tecniche di analisi per la strategia gerarchica agglomerativa

Le tecniche *agglomerative* (dette anche *aggregative* o *bottom-up*) sono le più utilizzate sia per la relativa semplicità con cui è possibile programmarle sia perché possono essere applicate ad un grande numero di elementi. Esse procedono partendo dalla situazione in cui ogni unità costituisce un gruppo a se stante; quindi si procede, con successive fusioni, all'aggregazione delle n unità seguendo un criterio di minimizzazione delle distanze (o massimizzazione delle somiglianze); in particolare, il procedimento prevede i seguenti passaggi:

a. nella matrice di prossimità si individua la distanza più piccola e si aggregano tra loro le unità più vicine;
b. si ricalcola la matrice delle distanze tenendo conto del gruppo ottenuto precedentemente, che sostituisce le unità aggregate;
c. si costituisce un nuovo gruppo sulla base della distanza più piccola trovata nella nuova matrice.

Il procedimento iterativo termina dopo $n-1$ passaggi quando si forma il gruppo che comprende tutte le unità. Al termine è possibile ricostruire l'intero procedimento rappresentandolo come un albero.

Il procedimento viene ripetuto $n-1$ volte, considerando distanze tra unità, tra unità e gruppi e tra gruppi, fino al punto in cui tutte le unità saranno confluite in un unico gruppo. Alla fine ogni partizione risulterà contenuta nella precedente.

Gli algoritmi utilizzati soddisfano un numero piccolo di proprietà statistico-matematiche ma possono essere applicati ad un numero molto grande di unità. Secondo molti autori le tecniche agglomerative utilizzate in analisi che riguardano sistemi omogenei (analisi ecologiche o di comunità) non conducono a risultati convincenti.

Le principali tecniche di aggregazione, molte delle quali possono essere applicate a matrici di distanze non-metriche[7] sono:

– Legame singolo (*single linkage*, *nearest neighbour*, *minimum method*, *analisi gerarchica singola di Johnson*, *Johnson min*): secondo questa tecnica la distanza tra due *cluster* è determinata sulla base della distanza tra i due elementi, appartenenti a due *cluster* diversi, più vicini; in altre parole i *cluster* vengono aggregati solo sulla base delle informazioni di due singoli elementi che risultano essere molto vicini. Tale tipo di procedimento, che tende a favorire l'aggregazione di due gruppi con unità vicine, produce un concatenamento tra le entità; per questo motivo tale approccio è particolarmente adatto ai casi in cui i *cluster* sono omogenei. Nell'albero prodotto da questa tecnica la distanza tra i due punti più estremi è minima tra tutte le rappresentazioni possibili con $n(n-1)/2$ distanze (*minimum spanning tree*). Tale metodo può essere applicato con qualsiasi misura di prossimità.

[7] Si ricordi che le distanze definite *non metriche* sono quelle che non soddisfano la disuguaglianza triangolare.

- Centroide (*pair-group centroid*, *group means*): il procedimento è analogo a quello utilizzato con il legame singolo, salvo che per il criterio di aggregazione che, in questo caso, considera la distanza quella esistente tra i centroidi dei due *cluster*. Il centroide è definiti come il punto medio dello spazio multidimensionale (centro di gravità del *cluster*). Rappresenta una tecnica più robusta di altre ma è molto sensibile alla presenza di *outlier*. Una variante di questo approccio (detta *weightes*) prevede che nel calcolo si utilizzi come peso la differenza delle dimensioni dei *cluster*; tale tecnica risulta così utile nei casi in cui vi siano considerevoli differenze nelle dimensioni dei *cluster*.
- Legame completo (*complete linkage*, *furthest neighbour, maximum method*, *analisi gerarchica completa di Johnson*, *Johnson max*): secondo questa tecnica la distanza tra *cluster* è determinata sulla base della distanza tra i due elementi, appartenenti a due *cluster* diversi, più lontani. Questa tecnica tende a produrre *cluster* armonici e compatti (a meno che non vi siano valori *outlier*), con una notevole omogeneità interna. È consigliabile applicare tale tecnica nei casi in cui gli elementi formano realmente blocchi naturali e distinti. Con questa tecnica può essere utilizzata qualsiasi misura di prossimità.
- Legame medio (*average linkage*, *pair-group average*, *group average*): secondo questa tecnica, la distanza tra due *cluster* è calcolata come la distanza media tra tutte le coppie di elementi appartenenti ai due diversi *cluster*; la tecnica tende ad unire *cluster* con piccole varianze. Una variante di questo approccio (detta *weighted*) prevede che nel calcolo si utilizzi la dimensione di *cluster* come peso; tale tecnica risulta così utile nei casi in cui vi siano considerevoli differenze nelle dimensioni dei *cluster*.
- Varianza minima di *Ward* (*criterio dell'inerzia*): tale tecnica si distingue da tutte le altre in quanto per valutare la distanza tra *cluster* utilizza l'analisi della varianza; in particolare essa cerca di minimizzare la varianza *within* tra due *cluster* che possono essere formati in ciascun passaggio. Si ricordi che la varianza (detta in questo caso *inerzia*) in un gruppo di elementi è pari alla media dei quadrati delle distanze dal centro di gravità del gruppo.
  Il criterio scelto per l'aggregazione di due gruppi è quello secondo il quale l'aggregazione porta ad un aumento minimo della varianza all'interno del gruppo[8].

[8] La formalizzazione dell'algoritmo di Ward è la seguente:
– la matrice delle distanze ***D*** tra le *n* unità di partenza è sostituita dalla matrice δ con:

$$\delta_{ij} = \left[\frac{(p_i p_j)}{(p_i + p_j)} \right] * d^2(e_i, e_j)$$

In pratica per tutti gli *i* e *j*;
– si cercano le due unità per le quali δ (indice del livello di aggregazione) è minore aggregandole in una classe di peso p_i+p_j;
– si calcolano le distanze tra le altre unità ed il gruppo precedentemente ottenuto;
– si cercano gli elementi (gruppi o unità più vicini, si aggregano in gruppi,
e così via.

La tecnica viene considerata molto efficiente e tende a produrre *cluster* di piccole dimensioni. Si applica prevalentemente su distanze euclidee ma può essere utilizzata anche con altri tipi di distanze.

– Metodo di densità (*density, k-linkage, density-seeking mode analysis*): tale tecnica si riferisce in realtà ad una classe di tecniche che utilizza stime di *densità di probabilità non parametriche*. Essa è stata sviluppata in una forma avanzata (detta *a due stadi*) nell'istituto che predispone il *package* statistico SAS.

Un'altra tecnica, detta *EML* e simile alla *varianza minima di Ward*, aggrega *cluster* per massimizzare la verosimiglianza a ciascun livello della gerarchia. L'esperienza pratica ha indicato che tale tecnica (studiata all'interno dell'istituto che predispone il *package* SAS) tende a produrre *cluster* di dimensione diversa (diversamente dalla tecnica di Ward).

Nel caso di prossimità misurate su scala discreta esiste la possibilità di utilizzare un'altra tecnica detta della somiglianza di McQuitty.

Nella scelta tra le diverse tecniche è importante tenere conto che:

– se la matrice è composta da distanze euclidee, sono consigliabili i metodi della media ponderata e quello di Ward;
– se l'obiettivo è quello di individuare gruppi omogenei al loro interno, indipendentemente dalla misura di prossimità utilizzata, è opportuno applicare il metodo del legame completo;
– se si ipotizzano gruppi non sferici, un certo concatenamento tra le unità o la presenza di dati anomali, è consigliabile applicare il metodo del legame singolo;
– se non si dispone di alcuna informazione sulla struttura (in termini di distanze) e sulla forma che dovrebbero avere i gruppi, è consigliabile applicare il metodo del legame singolo in quanto produce grappoli sicuramente ben definiti e separati e identifica *cluster* di qualsiasi forma.

Data la soggettività nella scelta del tipo di procedure, si consiglia di applicare più algoritmi e di confrontare i risultati ottenuti in modo da verificare se le unità si prestano ad essere classificate.

Il confronto tra le diverse tecniche aggregative non è semplice e può basarsi solamente su verifiche empiriche. Quindi per verificare le diversità e le analogie tra le tecniche nella capacità di rilevare la situazione reale, è necessario individuare una situazione concreta e chiaramente caratterizzata[9]. La possibilità di effettuare confronti è complicata dal fatto che i diversi algoritmi sono stati ideati con riferimento a logiche diverse. A tale proposito è stato definito il concetto di *partizione ben strutturata* secondo la quale due elementi appartenenti ad uno stesso gruppo devono avere sempre una distanza inferiore a quella di due elementi appartenenti a gruppi diversi. Una partizione ben strutturata si dirà *minimale* se conterrà il minor numero di gruppi. È dimostrabile che i metodi del legame singolo, del legame completo e del legame medio soddisfano le proprietà di *partizione ben strutturata minimale*.

[9] A tale proposito si ricordi che, in altri ambiti della statistica, esiste la possibilità di fare confronti metodologicamente più agevoli; si pensi alle ricerche sulla *robustezza* dei test che sono facilitate dalla conoscenza della distribuzione campionaria.

Indipendentemente dalla tecnica prescelta, la soluzione gerarchica presenta comunque due vantaggi:
– dà una visione completa delle partizioni ottenute nelle diverse iterazioni in termini di distanze;
– non richiede che venga definito in partenza il numero di *cluster*.

Il dendrogramma

Come si è detto, l'approccio gerarchico in partenza considera le *n* unità statistiche e, sulla base della matrice di prossimità tra le *n* unità, individua le due più vicine aggregandole. Il procedimento viene iterato per le successive aggregazioni che vengono effettuate sulla base della tecnica prescelta; al termine delle iterazioni tutte le unità vanno a comporre un unico gruppo. I risultati del procedimento iterativo di aggregazione possono essere rappresentati graficamente attraverso un albero detto anche *dendrogramma*. In esso gli oggetti sono rappresentati come nodi mentre la lunghezza del ramo indica la distanza tra i sottogruppi che vengono uniti. Una volta rappresentato, il dendrogramma consente di interpretare facilmente i risultati e di individuare l'iterazione che ha ottenuto la partizione ottimale, tenendo conto sia della scala di distanze tra i gruppi ottenuti che degli obiettivi dell'analisi. Un dendrogramma che chiaramente differenzia i gruppi di oggetti in genere presenta piccole distanze nei primi rami e grandi distanze negli ultimi rami.

Nella figura I. 1.3 è riportato un esempio di dendrogramma ottenuto a partire dalla matrice delle distanze aeree tra alcune città; la scala delle distanze riportata consente di valutare la distanza tra i raggruppamenti ottenuti; il metodo di aggregazione adottato è quello del legame singolo.

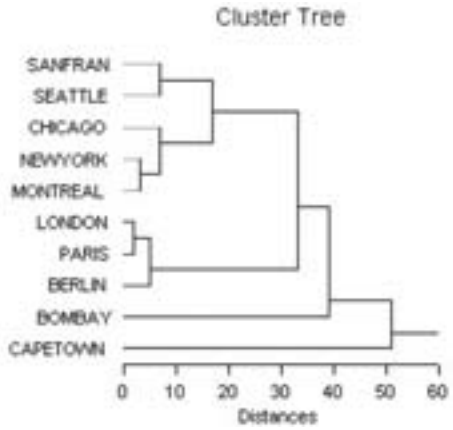

Fig. I. 1.3 Esempio di dendrogramma

Osservando tale dendrogramma, e tenendo conto della scala delle distanze, è facile individuare la partizione più facilmente interpretabile che è quella che ha raggruppato le città per continenti e collocazione rispetto alla longitudine (cinque gruppi).

Una volta accertato il livello di aggregazione ottimale, si individua il valore di distanza corrispondente dal quale si traccia una perpendicolare. I raggruppamenti che risultano alla sinistra di tale retta sono quelli che possono essere presi in considerazione sulla base delle considerazioni precedentemente fatte.

Two-way joining

I raggruppamenti discussi riguardavano genericamente *casi* (righe della matrice dei dati); l'applicazione dei metodi di *cluster* può avere senso anche se riguarda le variabili (colonne della matrice dei dati). In molti casi è anche possibile considerare entrambe le dimensioni di raggruppamento. In altre parole, l'interesse può essere rivolto anche al raggruppamento simultaneo di casi e variabili. A tale proposito è possibile immaginare uno studio medico in cui si siano raccolti dati rispetto a diversi indicatori di salute fisica (variabili) su un campione di malati di cuore (casi). In questo caso il ricercatore può essere interessato a identificare *cluster* di pazienti che sono simili rispetto a particolari *cluster* di misure simili di salute fisica[10].

Tale approccio all'analisi risulta comunque di difficile interpretazione. Comunque essa è considerata come un metodo che offre un potente strumento di analisi esplorativa dei dati.

1.2.2.2 Interpretazione di una soluzione gerarchica

Dopo aver individuato i gruppi si procede all'interpretazione della soluzione ottenuta per la quale è necessario verificare che i risultati siano coerenti con i dati e identificare le caratteristiche più importanti dei gruppi trovati.

L'interpretazione dei risultati prodotti dalle tecniche gerarchiche si presenta piuttosto complessa; essa dipende molto dagli elementi a disposizione e dagli scopi dell'applicazione dell'analisi dei *cluster*. In genere l'interpretazione si basa sul diagramma ad albero. La lettura di tale struttura può avvenire in modo

– *verticale*, in modo da osservare come si raggruppano i dati;
– *orizzontale*, in modo da vedere ad un determinato livello quali elementi si raggruppano tra loro.

Tale rappresentazione presenta però una distorsione causata sia dalla misura di distanza/somiglianza tra casi utilizzata che dal metodo di raggruppamento adottato.

Per valutare la validità e la bontà della partizione ottenuta e generata dalla tecnica scelta si procede confrontando la configurazione ottenuta con i dati originali di prossimità. In particolare si confrontano:

– le misure di prossimità originarie,
– le misure cofenetiche; queste rappresentano i valori di dissomiglianza, deducibili dal dendrogramma, che due osservazioni hanno nel momento in cui sono state combinate in un *cluster*; più tali valori sono elevati più gli elementi aggregati sono eterogeni.

[10] In questi casi è molto importante procedere alla standardizzazione delle misure da sottoporre ad analisi.

Se la partizione è valida, la relazione tra tali misure dovrebbe essere molto stretta. Tale relazione viene misurata attraverso il coefficiente di correlazione cofenetico che consente di valutare la distorsione del diagramma ad albero e che quindi può essere considerato come misura della concordanza tra la soluzione ottenuta e la matrice di prossimità iniziale:

$$R_c = \frac{\sum_{i>j}\left(d_{ij} - m\right) * \left(d_{ij}^* - m^*\right)}{\sqrt{\sum_{i>j}\left(d_{ij} - m\right)^2 * \sum_{i>j}\left(d_{ij}^* - m^*\right)^2}}$$

dove

d_{ij} prossimità originaria tra i casi i e j
d_{ij}^* distanza cofenetica tra i casi i e j
m media delle prossimità
m^* media delle distanze *cofenetiche*

Tale indice ha la proprietà di diminuire al crescere della distorsione.

La soluzione è considerata di alta qualità e il dendrogramma rappresenta un'appropriata sintesi dei dati se il valore del coefficiente è vicino a 1. In caso contrario il dendrogramma può essere visto come una semplice descrizione dell'output dell'algoritmo di aggregazione adottato. Tale misura può essere utilizzata anche per confrontare soluzioni alternative ottenute utilizzando tecniche diverse.

Per misurare la relazione tra prossimità di partenza e distanze *cofenetiche* è possibile utilizzare anche un altro indice basato sulla *distanza di Minkowski* (v. volume 1):

$$D_c = \frac{\left(\sum_{i>j}\left|d_{ij} - d_{ij}^*\right|^r\right)^{1/\lambda}}{\left(\sum_{i>j} d_{ij}^{\lambda}\right)^{1/\lambda}}$$

dove, come si sa, a seconda del valore assunto dal parametro λ, la distanza calcolata assume un modello diverso; se

– $\lambda=1$ la distanza calcolata è uguale a quella assoluta (*city block*),
– $\lambda=2$ la distanza calcolata è uguale a quella euclidea.

La complessità dell'interpretazione dei risultati ottenuti attraverso l'approccio gerarchico è compensata da alcuni pregi rappresentati principalmente dal fatto che non richiedono di prefissare in anticipo il numero dei *cluster* cui si vuole giungere e possono essere utilizzati in genere sia rispetto alle variabili sia rispetto alle osservazioni.

1.2.2.3 Tecniche di analisi per la strategia non gerarchica

I metodi non gerarchici hanno l'obiettivo di aggregare in un'unica soluzione le unità in r gruppi in modo tale che le unità che sono all'interno dello stesso gruppo siano più omogenee possibile, mentre i gruppi siano tra loro più disomogenei pos-

sibile. Per poter ottenere la soluzione è necessario che il ricercatore faccia un'ipotesi riguardante il numero di *cluster* presenti tra gli elementi osservati.

Procedimento

Importante per l'applicazione dell'approccio non gerarchico è la definizione di un criterio che stabilisce la *qualità della partizione*. La scelta del criterio deve tener conto di aspetti qualitativi quali la conoscenza dell'insieme studiato e delle finalità della ricerca.

Un esempio di criterio può essere quello che definisce la ripartizione migliore come quella che presenta la massima distanza tra i centroidi dei gruppi e la minima tra le unità interne ai gruppi; uno dei metodi più utilizzati è quello che verifica la media dei gruppi; questo è il motivo per cui l'analisi non gerarchica è detta anche *k-means cluster analysis*. Il criterio più utilizzato è comunque quello che fa dipendere l'assegnazione degli elementi ai *cluster* da uno dei seguenti l'obiettivi:

a. minimizzare la variabilità all'interno dei *cluster,*
b. massimizzare la variabilità tra *cluster*.

Stabilito il criterio si procede secondo un procedimento iterativo che richiede l'adozione di una soluzione approssimata; tale soluzione può essere individuata attraverso diverse tecniche tra le quali è possibile citare le seguenti:

– *limite sequenziale* (*sequential threshold method*): individuato casualmente un valore (detto *seme*) considerato centroide, tutti gli oggetti che rientrano all'interno di uno specificato valore limite da tale centro sono considerati appartenenti allo stesso gruppo; dopo aver selezionato un nuovo centroide il procedimento viene ripetuto per i punti che non hanno trovato collocazione in precedenza.
– *Limite parallelo* (*parallel threshold method*): vengono definiti simultaneamente molti centroidi; gli oggetti che rientrano in uno dei definiti limiti sono raggruppati al centro più vicino.
– *Ottimizzazione della suddivisione* (*optimizing partitioning method*): gli elementi possono essere successivamente riassegnati ad altri *cluster* al fine di ottimizzare un criterio generale, per esempio stabilendo una media delle distanze all'interno di un *cluster*.
– *Centri mobili*: scelte casualmente *r* unità assunte come centri dei *r* gruppi ipotetici, le altre unità vengono associate al centro meno distante (*prima partizione*). Si calcolano i centroidi reali degli *r* gruppi ottenuti. Quindi si procede all'assegnazione di ciascuna delle *n* unità al centroide più vicino (*seconda partizione*), si ricalcolano i centroidi e si ripete il procedimento fino a quando due iterazioni successive non producono partizioni identiche. Con un grande numero di unità, difficilmente viene soddisfatto il criterio entro un numero ragionevole di iterazioni; per questo motivo è possibile fissare a priori il numero di iterazioni. Quando viene raggiunto il numero stabilito di iterazioni o viene soddisfatto il criterio, la procedura si ferma e l'ultima partizione ottenuta viene adottata.
– *Nuvole dinamiche*: per migliorare la classificazione, è stato proposto un procedimento che prevede, per l'individuazione della prima soluzione approssimata, la selezione casuale per ciascun gruppo non di un'unica unità per ciascun gruppo ma di un insieme di unità (metodo delle *nuvole dinamiche*).

– *Raggruppamenti stabili*: vengono definite molte soluzioni che confrontate consentiranno di individuare le unità che risultano essere sempre allocate nello stesso gruppo.

Il difetto che accomuna tutte le tecniche è quello di essere molto influenzate dalla presenza di dati anomali.

Rispetto alle soluzioni gerarchiche, i risultati prodotti dall'approccio non gerarchico sono sicuramente più semplici da interpretare in quanto rappresentati da un'unica partizione e il ricercatore non deve fare altro che constatare la suddivisione prodotta e procedere all'interpretazione dei risultati sulla base delle proprie ipotesi.

Determinazione del numero ottimale di gruppi

Uno dei principali problemi nell'applicazione dell'approccio non gerarchico è dato dal fatto che richiede la specificazione del numero dei gruppi da individuare. Non sempre però è possibile determinare tale numero. A tal fine può essere utile procedere in uno dei seguenti modi.

– Rappresentazione grafica di diverse soluzioni: la rappresentazione grafica dei risultati di diverse analisi con un numero variabile di gruppi consente di identificare la soluzione che produce una maggiore discontinuità tra i gruppi identificati e una maggiore omogeneità all'interno dei gruppi. Si tratta di una procedura che però non sempre dà risultati illuminanti.
– Verifica statistica della bontà della soluzione: sulla soluzione ottenuta è in molti casi possibile applicare test statistici per verificare se essa è significativamente diversa da una ottenibile per caso. In genere si verifica se la distanza tra le medie dei due gruppi è significativa. Per stabilire la significatività dell'applicazione delle tecniche che producono partizioni ottimizzando le funzioni della matrice di devianze-codevianze, è possibile utilizzare la statistica *lambda* di Wilks. Si ricordi, comunque, che, con n grande, è molto più facile che le differenze tra i gruppi ottenuti risultino statisticamente significative.
– Analisi gerarchica: su un gruppo di dati l'analisi non gerarchica può essere preceduta da una analisi gerarchica; l'analisi dei risultati ottenuti consente di fare una prima ipotesi sul numero di gruppi identificabili

Esistono alcune statistiche sintetiche che consentono di decidere oggettivamente il numero dei gruppi. Tra queste si ricordano C di Calinski e Harabasz e M di Marriot.

Infine, è opportuno ricordare che per verificare il numero ottimale di gruppi è anche possibile applicare l'analisi discriminante (trattata più avanti) successivamente all'individuazione di una soluzione.

1.2.3 Particolari questioni

Verifica della significatività statistica

Una delle critiche che in genere si muovono alla *cluster analysis* è quella di giungere a soluzioni indeterminate, soggette a decisioni arbitrarie relative alle informazioni iniziali, alle tecniche di raggruppamento, all'interpretazione soggettiva dei risultati, non sottoponibili a verifica statistica.

Diversamente da altre procedure statistiche, la *cluster analysis* spesso viene utilizzata quando non si hanno ipotesi a priori o quando si è nella fase esplorativa di un'analisi. Dato che tale analisi si pone l'obiettivo di ricercare la soluzione più significativa possibile, la verifica della significatività non si presenta veramente importante. Importante è invece che l'applicazione della *cluster analysis*, pur rientrando tra i metodi di analisi essenzialmente esplorativa, sia preceduta e accompagnata dalla definizione di modelli interpretativi.

Confronto tra le tecniche di *clustering*

Sapere che le diverse tecniche di *clustering* possono produrre risultati anche molto diversi tra loro non è solo una curiosità accademica. È molto importante conoscere la forza e la debolezza delle diverse tecniche ed esplorare i motivi delle differenze prima di procedere all'analisi. Si è già visto come certe tecniche presentano dei *bias* da tener presenti (per esempio, il metodo del legame singolo tende a produrre *cluster* allungati e incatenati).

Un modo per valutare le differenze tra le diverse tecniche di *clustering* è quello di valutare in quale misura riproducono la struttura presente e conosciuta dei dati. Tali valutazioni, effettuate in genere su dati simulati, spesso sono difficili da interpretare e possono risultare contraddittori.

I fattori che sembrano influenzare maggiormente i risultati di tali analisi sono:

– gli elementi che definiscono la struttura di *cluster* ovvero forma, dimensione (assoluta e relativa) e numero di *cluster*;
– la presenza di *outlier*;
– il livello di sovrapposizione tra *cluster* espresso in termini di spazio occupato da due o più *cluster* (presenza di *cluster* ben separati, adiacenti o sovrapposti);
– il tipo di misura di somiglianza/distanza prescelto.

La gestione degli *outlier*

Nell'ambito dell'analisi dei gruppi, gli *outlier* possono essere definiti come casi che presentano particolari valori in contrasto con quelli delle altre osservazioni; in altre parole, si tratta di casi che presentano una combinazione di valori unica, identificabile in modo distinto da quelle delle altre osservazioni. In questo senso un *outlier* potrebbe definire un *cluster* autonomo; ciò potrebbe però condurre ad una soluzione finale con numero di *cluster* inaccettabile e non parsimonioso. Sta al ricercatore valutare se tali casi presentano comunque valori legittimi o informativi che possono essere considerati nel contesto di analisi. In un'analisi su dati campionari, gli *outlier* potrebbero essere indicativi di caratteristiche della popolazione che non sarebbero emerse in altri momenti dell'analisi. Và però tenuto presente che la presenza di *outlier* comunque porta ad una distorsione anche piuttosto seria dei risultati. Per questi motivi è importante che il ricercatore esamini con attenzione tali valori e valuti la loro influenza.

La questione diventa particolarmente seria nell'ambito delle strategie agglomerative quando il caso estremo potrebbe entrare a far parte del gruppo rispetto al quale registra il più alto livello di somiglianza anche se in realtà tale livello, se confrontato con gli altri, molto basso.

1.3 Altri approcci di analisi per l'individuazione di gruppi: gli alberi di classificazione

L'approccio detto *trees* si configura come metodo asimmetrico; esso utilizza la lunghezza dei rami di un albero per rappresentare la distanza tra gli oggetti; ammettendo la variazione delle distanze tra *cluste*r, la tecnica produce un diagramma ad albero che presenta e che inizia con un nodo che si dirama in molti rami di diversa lunghezza. Gli oggetti all'interno di un *cluster* possono essere così confrontati mettendo in luce la distanza orizzontale lungo i rami che li collegano.

In origine tale approccio, utilizzato soprattutto a fini predittivi, proponeva procedure e algoritmi per l'identificazione automatica delle interazioni tra variabili; attualmente i metodi detti *tree-fitting* sono considerati una alternativa all'analisi dei cluster.

È possibile distinguere due differenti approcci: *classification tree* e *regression tree*. Nel primo la variabile dipendente è categorica, mentre nel secondo la variabile dipendente può essere continua.

L'identificazione dell'albero prende avvio da un unico nodo che contiene l'intero gruppo di casi da raggruppare. Ciascuno dei successivi nodi identificati contiene un sottogruppo dei casi appartenenti al nodo precedente. Quindi, ciascun nodo contiene la somma dei gruppi appartenenti ai nodi collegati ad esso ed immediatamente successivi. Ciascun nodo può essere immaginato come un cluster di oggetti/casi che viene successivamente suddiviso nei successivi rami dell'albero. La rappresentazione ad albero è molto simile alla struttura dei dendrogrammi visti in precedenza; la differenza sta nel fatto che, nei modelli predittivi, i nodi e i rami sono identificati dai valori della variabile indipendente e dipendente.