CESIT - Centro Studi sui Sistemi di Trasporto "Carlo Mario Guerci"
Piazza Bovio 14 80133 Napoli
Working paper series n. 1-2015

–

LA CLASSIFICAZIONE AUTOMATICA PER UNO STUDIO DEL SISTEMA DEI TRASPORTI

Rosaria Lombardo

LA CASSIFICAZIONE AUTOMATICA

CLASSIFICAZIONE E CLUSTERING

1. Introduzione

L'analisi dei gruppi, o cluster analysis, è un insieme di tecniche atte a formare dei gruppi"omogenei", secondo un certo criterio.

I metodi di **classificazione automatica** (CA) hanno come obiettivo il raggruppamento delle unità della tabella unità-variabili per mezzo di algoritmi formalizzati e costruiti in base a criteri di ottimizzazione predefiniti.

Tra le unità si ipotizza l'esistenza a priori di raggruppamenti o se ne richiede la determinazione: un metodo di classificazione può o confermare gruppi che già costituiscono una realtà concreta o individuare classi che risultino omogenee rispetto agli indicatori individuati e al metodo applicato. La costruzione dei cluster si può effettuare in molti modi, in funzione sia della scelta del criterio di "misura della somiglianza" (o della "differenza") tra le unità, sia delle diverse strategie di raggruppamento (gerarchiche o non-gerarchiche; divisive o agglomerative).Ciascun algoritmo di raggruppamento individuerà dei gruppi in funzione dei parametri che specificano il metodo e delle variabili considerate. Ogni scelta tra questi criteri porta, in genere, a classificazioni differenti. Questo significa che in una classificazione due dati apparterranno nello stesso gruppo mentre apparterranno a gruppi diversi per un altra classificazione.

Qualunque sia il metodo di classificazione e la misura di somiglianza-dissimiglianza scelta, la cluster analisi procede poi “automaticamente” a raggruppare i dati. La sola scelta che resta, una volta messo in moto l’algoritmo di formazione dei cluster, è quando fermarsi, ovvero quando ci si ritiene “soddisfatti” della classificazione ottenuta.

E’ quindi di fondamentale importanza capire il significato dei vari elementi costitutivi deglialgoritmi di formazione dei cluster, per decidere quale, tra i tanti possibili criteri è quello piùindicato ai dati che vogliamo analizzare.

Classificare significa distinguere e separare ciò che appare diverso ed unire in gruppi omogenei di ciò che invece si presenta simile; queste semplici operazioni consentono di organizzare collezioni anche molto vaste di informazioni.

La classificazione è uno dei processi fondamentali della scienza e costituisce parte integrante di ogni processo di apprendimento, mediante la classificazione è possibile individuare la struttura intrinseca delle informazioni analizzate e le relazioni che legano le stesse; ottenendo una migliore e più completa comprensione del problema analizzato.

Oltre ad agevolare la conoscenza, la classificazione consente un’adeguata organizzazione delle informazioni; uno schema di classificazione infatti può rappresentare un metodo estremamente conveniente per organizzare un vasto insieme di dati in modo che il recupero delle informazioni possa essere effettuato in maniera efficiente.

La memorizzazione ed il successivo reperimento delle informazioni risultano molto semplici e rapidi, poiché la classificazione permette di ridurre la mole di informazioni da gestire, rendendo sufficiente analizzare le sole classi od un loro elemento rappresentativo piuttosto che l’intero insieme dei dati, inoltre la migliore conoscenza della struttura intrinseca contribuisce ad una più efficace procedura di ricerca.

La reale importanza della classificazione appare evidente se si considera il fatto che la comprensione della struttura intrinseca dell’informazione, la sua organizzazione ed il suo recupero efficiente rappresentano aspetti di fondamentale importanza in svariate discipline scientifiche. Ad esempio la segmentazione del mercato rispetto alle diverse esigenze e preferenze dei potenziali consumatori rappresenta solo uno dei motivi per cui la classificazione sia divenuta elemento imprescindibile della scienza, in particolare di quelle economiche.

Aristotele (384 - 322 a.C.) fu il primo a dare un significato scientifico al concetto di classificazione introducendo la logica “classificatoria”, ma ancora oggi dopo più di duemila anni non esiste una

vera e propria “scienza della classificazione”, nonostante tale attività sia così importante e così largamente utilizzata.

La sempre più rapida crescita del volume di informazioni a disposizione dell’uomo e della scienza, ha condotto alla ricerca ed allo sviluppo di strumenti nuovi e più efficienti per organizzare ed analizzare le informazioni.

2. Cluster Analysis

La parola inglese *cluster* è difficilmente traducibile, letteralmente significa “grappolo”, ma anche “ammasso”, “sciame”, “agglomerato”, parole che visivamente richiamano alla mente una o più entità costituite da elementi più piccoli, omogenei tra loro ma allo stesso tempo distinti da altri elementi esterni al cluster stesso.

Una definizione di cluster è peraltro difficile da trovare anche in letteratura, normalmente si parla di coesione interna ed isolamento esterno [Everitt, 1980], oppure di elevato grado di associazione naturale tra gli elementi di un cluster e di relativa distinzione tra cluster differenti [Anderberg, 1973], in modo più generale si può definire cluster una parte dei dati (un sottoinsieme della popolazione in analisi) che consiste di elementi molto simili rispetto alla rimanente parte dei dati [Mirkin, 1996]. La generalità delle definizioni di cui sopra è determinata principalmente dalla assoluta impossibilità di formalizzare matematicamente il concetto di cluster in maniera univoca, del resto empiricamente si possono osservare molti tipi differenti di cluster: sferici, ellissoidali, lineari.

Le tecniche che permettono di ottenere i cluster dai dati osservati sono molteplici ed utilizzano procedure differenti, comunemente vengono denominati *algoritmi di clustering* per sottolineare il carattere automatico di tale procedura.

E’ possibile far risalire la nascita di tale disciplina all’inizio del secolo scorso, quando si cominciò ad avvertire la necessità di sostituire l’occhio ed il cervello umano nell’attività classificatoria con strumenti maggiormente precisi, in grado di gestire enormi quantità di informazioni e non limitati a spazi tridimensionali.

Solamente con lo sviluppo e la diffusione dei moderni computer le procedure di clustering cominciarono ad essere impiegate con successo in numerose discipline: nelle scienze naturali, nell’ambito della *tassonomia numerica* [Sneath&Sokal, 1973], allo scopo di rendere più veloce e precisa l’opera di classificazione delle specie viventi; nelle scienze sociali per individuare gruppi

socio-economici e segmenti di mercato; nel campo dell’intelligenza artificiale ed in particolare nella *pattern recognition*, per l’analisi, il confronto ed il riconoscimento dei dati. Seguendo questo percorso, pur in assenza di un rigoroso fondamento teorico, ha avuto origine la disciplina oggi nota come *cluster analysis*.

Generalmente gli algoritmi di clustering cercano di separare un insieme di dati nei suoi cluster costituenti, evidenziando i *gruppi naturali*, ossia la struttura classificatoria relativa ai dati stessi, si suppone che i dati analizzati possiedano una propria classificazione e compito degli algoritmi di clustering è proprio trovare la struttura classificatoria che meglio si adatta alle osservazioni.

Il compito degli algoritmi di clustering è duplice, da un lato trovare la struttura classificatoria intrinseca dei dati, i cosiddetti *gruppi naturali*, dall’altro assegnare ogni elemento alla rispettiva classe di appartenenza, per fare questo generalmente gli algoritmi cercano di classificare le osservazioni in gruppi tali che il grado di *associazione naturale* sia alto tra i membri dello stesso gruppo e basso tra i membri di gruppi differenti. Si può quindi affermare che l’essenza della cluster analysis può anche essere vista come l’assegnazione di appropriati significati ai termini “gruppi naturali” ed “associazione naturale”.

3. Classificazione degli algoritmi di clustering

Data la varietà dei metodi e degli algoritmi di clustering disponibili è opportuno presentarne una classificazione[1] [Sneath&Sokal, 1973], che permetta di comprendere quali siano le principali caratteristiche delle diverse procedure.

[1] Quella presentata non è l’unica classificazione degli algoritmi di clustering, vedi [Bezdek, 1981]. Inoltre è opportuno precisare che sono stati esclusi dalla presente trattazione i metodi di clustering che utilizzano la teoria dei grafi.

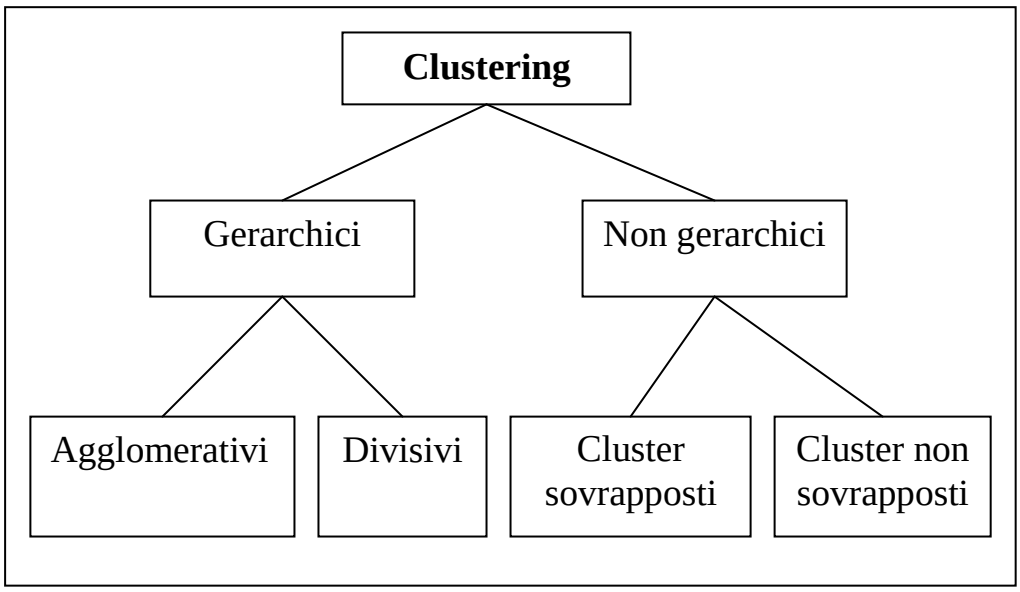


Figura 1 - Classificazione degli algoritmi di clustering

Metodi gerarchici e non gerarchici: costituisce la distinzione principale, si riferisce sia al metodo usato per ottenere i cluster che alla struttura del risultato dell'algoritmo stesso. I metodi gerarchici producono delle tipiche strutture ad albero, di tipo ricorsivo o annidato; i cluster dei livelli più alti sono aggregazioni di cluster dei livelli più bassi dell'albero. I metodi non gerarchici vengono anche definiti partitivi poiché dividono l'insieme dei dati in partizioni, le quali sono costituite solamente dai singoli elementi oggetto della classificazione, l'insieme dei dati da classificare viene quindi diviso in più sottoinsiemi o cluster senza ulteriori suddivisioni all'interno di ogni cluster.

Metodi agglomerativi e divisivi: si riferisce al metodo con il quale operano gli algoritmi gerarchici ed anche alla direzione seguita per costruire lo schema classificatorio (albero gerarchico). I metodi agglomerativi procedono dal basso verso l'alto unendo i singoli elementi mentre i metodi divisivi procedono dall'alto verso il basso scindendo i cluster in altri più piccoli.

Metodi con sovrapposizione e senza sovrapposizione (dei cluster): riguarda i cluster prodotti con i metodi non gerarchici o partitivi e si riferisce ai cluster ottenuti per mezzo dell'algoritmo. I cluster si definiscono non sovrapposti quando ogni elemento appartiene ad uno e ad un solo cluster; i cluster sovrapposti vengono anche denominati *fuzzy cluster* poiché ogni elemento può appartenere a più cluster differenti per mezzo di un *grado di appartenenza* definito come segue:

$$\mu_{c_i}(x_i) \in [0,1]$$

L’elemento x_i appartiene al cluster C_i con un grado di appartenenza compreso nell’intervallo [0,1]. Questa estensione della cluster analysis classica deriva dalla considerazione che nella realtà molti fenomeni non si presentano con contorni ben definiti, ed un elemento può appartenere contemporaneamente a più cluster differenti. Ad esempio nel caso si vogliano classificare diverse specie di frutta, gli ibridi si troveranno in una posizione intermedia tra i cluster dei frutti dal cui incrocio sono stati generati, ed inserirli forzatamente in uno solo dei due cluster rappresenta una errata interpretazione della realtà [Bezdek, 1981].

4. Struttura generale degli algoritmi di clustering

La necessità di elaborare un metodo algoritmico per la classificazione dei dati viene di solito spiegata in letteratura con un esempio alquanto convincente: se si pensasse di enumerare tutte le possibili strutture classificatorie di un insieme di *n* dati per confrontarle e quindi scegliere la migliore con un procedimento di tipo esaustivo, ci si troverebbe di fronte ad numero finito ma enorme di possibili schemi di classificazione tale da rendere computazionalmente intrattabile il problema; tale numero corrisponde alla seguente espressione:

$$\frac{1}{c!}\left[\sum_{j=1}^{c}\binom{c}{j}(-1)^{c-j} j^n\right]$$

Volendo quindi classificare venticinque osservazioni in dieci classi distinte (*n*=25, *c*=10) si dovrebbero esaminare 10^{18} strutture differenti, una mole di dati troppo grande da trattare persino per i più potenti computer. Il metodo su cui si fondano gli algoritmi gerarchici agglomerativi si basa su una procedura iterativa di unione degli elementi più simili due alla volta, il processo si ripete quindi *n*−*c* volte (15 volte per l’esempio considerato sopra).

I metodi partitivi invece utilizzano una procedura iterativa nota con il nome di *ciclo di Picard* [Bezdek, 1981], che comprende una fase iniziale in cui viene definita,generalmente in modo casuale, una struttura iniziale, che viene successivamente “aggiornata” ad ogni iterazione fino alla convergenza ad una struttura stabile. Questo metodo non permette di conoscere il numero esatto di iterazioni che dovranno essere effettuate; per cautelarsi da tale inconveniente molti algoritmi, oltre

al criterio di convergenza appena illustrato, utilizzano un criterio aggiuntivo, ossia terminano in ogni caso dopo un numero predeterminato di iterazioni.

In ogni caso tutti i metodi partitivi convergono in un numero solitamente non molto elevato di iterazioni, in generale, tanto più velocemente quanto più la struttura iniziale si avvicina alla reale struttura classificatoria. La fase di inizializzazione costituisce quindi un aspetto molto importante dell'algoritmo.

Nonostante la varietà degli algoritmi di clustering e la diversa metodologia utilizzata per trovare i gruppi naturali dei dati, è possibile individuare alcuni punti comuni a tutti gli algoritmi. Ovviamente esistono numerose differenze tra i diversi metodi, in special modo confrontando metodi gerarchici e non gerarchici; tuttavia entrambe le metodologie partono dalla medesima premessa: raggruppare gli elementi simili e separare quelli diversi, è quindi opportuno illustrare sommariamente gli aspetti che caratterizzano questi algoritmi.

In primo luogo per trovare la classificazione ottimale dei dati è necessario poter misurare la diversità o la somiglianza tra i diversi elementi, serve quindi uno strumento che traduca numericamente tali concetti. La letteratura presenta due differenti soluzioni relativamente alle misure di similarità o dissimilarità: un primo approccio che utilizza le misure di distanza ed un secondo che utilizza i coefficienti di correlazione quali misure di similarità tra elementi.

L'utilizzo di tali funzioni si differenzia fortemente tra metodi gerarchici e non gerarchici; i primi (nel caso di algoritmi agglomerativi) utilizzano la matrice delle distanze, che risulta quadrata, simmetrica ed in cui ogni elemento rappresenta la distanza tra due osservazioni.

Ad ogni iterazione la matrice delle distanze viene aggiornata, tenendo conto che gli elementi da analizzare sono diminuiti di un'unità, per effetto della fusione nel caso dei metodi agglomerativi; oppure sono aumentati di un'unità nel caso di algoritmi divisivi.

I metodi non gerarchici invece non si servono della matrice delle distanze, ma ad ogni iterazione la struttura viene aggiornata calcolando la distanza degli n elementi rispetto ai c centroidi dei cluster, vengono quindi calcolate nc distanze ad ogni passo[2].

Oltre ad una misura di distanza è necessaria una struttura che consenta di rappresentare la classificazione dei dati: gli algoritmi gerarchici utilizzano i *dendrogrammi*, strutture gerarchiche che rappresentano le unioni (o le scissioni) che l'algoritmo esegue ad ogni iterazione; quelli partitivi

[2]L'algoritmo "fuzzy c-means" ne calcola nc^2 ad ogni iterazione.

utilizzano un vettore di indici o più comunemente la *matrice di partizione*, una matrice $c \times n$ dove ogni elemento indica l'appartenenza di ogni osservazione rispetto ad ogni cluster.

Alcuni metodi gerarchici e quasi tutti i metodi partitivi si servono di elementi prototipici nell'elaborazione della struttura classificatoria, in queste procedure il tipo di centroide (solitamente elemento medio o mediano del cluster) ed il loro calcolo nella procedura iterativa incidono negativamente sulla complessità computazionale dell'algoritmo, anche se questo accorgimento rende più semplice ed intuitiva la procedura per determinare la corretta partizione dei dati.

Non esiste un metodo di clustering universalmente applicabile, i risultati forniti dall'algoritmo dipendono fortemente da tutti gli elementi considerati; utilizzare un metodo gerarchico od uno partitivo, una funzione di distanza piuttosto che un'altra o un particolare tipo di centroide conducono all'identificazione di strutture differenti per il medesimo campione di osservazioni, tale "debolezza" ha generato dubbi e perplessità.

Alcuni ricercatori hanno effettuato studi comparativi tra le numerose metodologie disponibili, cercando di agevolare la scelta tra i diversi metodi da parte dell'utente, i confronti si basano sia sulla tipologia di dati da analizzare che sulle caratteristiche della struttura classificatoria [Anderberg, 1973].

Per poter confrontare i diversi algoritmi, verificare la bontà dei risultati e stabilire se quella ottenuta è una struttura classificatoria plausibile, la procedura di clustering viene arricchita da una fase finale di *validazione dei risultati*, nella quale si ricorre a strumenti come le *funzioni di validazione* per misurare alcune caratteristiche proprie della struttura ritenute desiderabili, ad esempio, si richiede che la varianza all'interno dei cluster sia minima mentre tra cluster differenti deve essere elevata.

5. Misure di Distanza e Spazi Metrici

Parte fondamentale di ogni algoritmo di clustering è un'appropriata misura di distanza o dissimilarità che permetta di tradurre numericamente i concetti, in precedenza esposti, di associazione tra elementi simili e distinzione tra elementi appartenenti a cluster diversi; tale necessità si presenta sia per i metodi gerarchici che per quelli partitivi.

E' opportuno ricordare che alcuni autori traducendo letteralmente il significato del termine *associazione naturale* tra elementi utilizzano nei loro algoritmi delle misure di *similarità* piuttosto che misure di *dissimilarità*, ma data la ovvia relazione di complementarità esistente tra le due misure è opportuno soffermarsi solo sulle seconde.

L’utilizzo delle misure di distanza negli algoritmi di clustering appare più chiaro se si immagina ogni elemento da classificare come un punto in uno spazio iperdimensionale, per semplicità è possibile riferirsi ad uno spazio bidimensionale come in figura 2, dove sono esemplificati due cluster ben distinti, la loro forma approssimativa ed i loro “centroidi” (le croci poste al centro delle linee tratteggiate che rappresentano il perimetro dei cluster).

In questa rappresentazione la struttura classificatoria dei dati è facilmente individuabile, la distanza tra i punti (elementi) appartenenti al medesimo cluster è inferiore a quella tra punti appartenenti a cluster differenti; si può ancora notare che i centroidi, che rappresentano i centri dei cluster, hanno delle distanze minime rispetto agli altri elementi del cluster; infatti la minimizzazione della distanza dei punti attorno al centroide del cluster è uno dei criteri più utilizzati negli algoritmi di clustering.

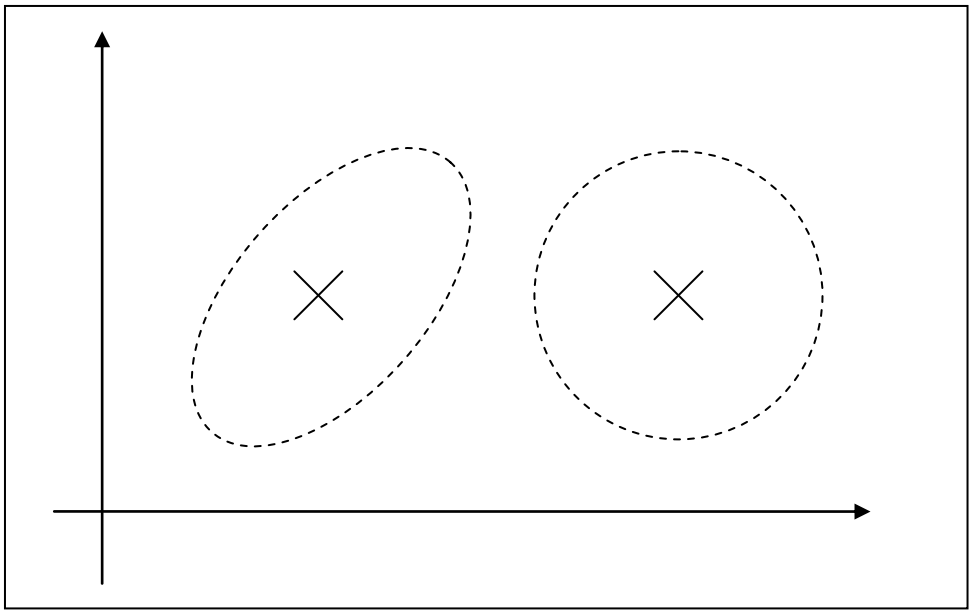


Figura 2 - Clustering in uno spazio bidimensionale

Questo ovviamente rappresenta un semplice esempio, utilizzato allo scopo di chiarire il motivo dell’utilizzo delle misure di distanza nella cluster analysis; generalmente, anche nei problemi più comuni, ogni elemento viene descritto da un vettore con più di tre elementi, anche in questo caso l’approccio rimane valido nonostante non sia effettuabile una verifica *ex post* basata su informazioni di tipo grafico.

Il primo elemento per la costruzione dell’algoritmo di raggruppamento è la “misura” che intendiamo adottare per valutare la “somiglianza” o la “dissomiglianza” tra due unità. La “somiglianza” deve tener conto dell’insieme di tutte le variabili.

La misurazione della “dissomiglianza” avviene attraverso la scelta di una funzione delle coppie di variabili misurate nei due casi. Questa funzione prende il nome generico di distanza.

Le scelte possibili sono molte, alcune adatte solo a dati di tipo numerico, altre utilizzabili sia per dati numerici che categoriali.

Nonostante la varietà delle scelte possibili, ci sono alcune caratteristiche comuni a tutte le possibili “distanze”.

A questo punto è necessario definire formalmente i concetti di misura di dissimilarità e di misura di distanza: sia S la rappresentazione simbolica di uno spazio di misura e siano **x**,**y**e **z** $\in S$ tre punti qualsiasi in S. Si definisce una misura di dissimilarità o “semimetrica” una funzione $d(x, y) : S \times S \Rightarrow R^{+}$ che soddisfa le seguenti condizioni:

1. d(**x**,**y**) = 0 se e solo se **x** = **y**,
2. d(**x**,**y**) ≥ 0 $\quad \forall x, y \in S$
3. d(**x**,**y**) =d(**y**,**x**) $\quad \forall x, y \in S$

La prima condizione indica la riflessività della relazione, la seconda richiede che la distanza, sia comunque non negativa, la terza indica infine la simmetria. Se oltre alle sopra elencate condizioni la funzione soddisfa anche la seguente, la funzione di distanza può essere definita una *metrica*:

4. d(**x**,**y**)≤d(**x**,**z**)+d(**y**,**z**) $\quad \forall x, y, z \in S$

Questa condizione, comunemente definita *disuguaglianza triangolare*, richiede che la distanza tra i punti **x** ed **y** sia minore od al più uguale alla somma delle distanze tra i due punti ed un terzo punto **z** distinto dai precedenti (vedi figura 3).

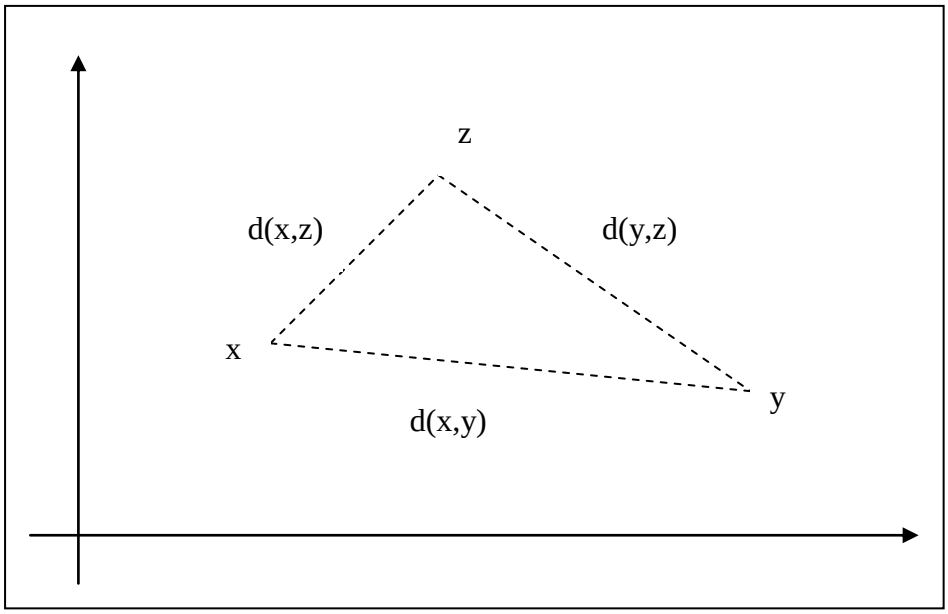


Figura 3 - Disuguaglianza triangolare

Se la funzione soddisfa la seguente, più forte versione della disuguaglianza triangolare la misura di distanza costituisce una *ultrametrica*:

5. $d(\mathbf{x},\mathbf{y}) \leq \max\{d(\mathbf{x},\mathbf{z}), d(\mathbf{y},\mathbf{z})\} \quad \forall x, y, z \in S$

La scelta tra indici di dissimilarità e metrica e' legata al tipo di dati che si hanno a disposizione. Per dati di tipo numerico (quantitativi) possiamo utilizzare delle misure di distanza, ovvero delle metriche. Per dati di tipo qualitativo bisogna utilizzare misure matching-type, cioè di associazione (similarità o dissimilarità).

La scelta della misura di dissimilarità è legata alla natura dei dati osservati, risultando infatti diverse le misure da adottare a seconda che gli oggetti siano descritti da variabili numeriche, da frequenze o da variabili nominali. Inoltre, mentre un indice di dissimilarità consente il solo confronto tra le caratteristiche di coppie di elementi dell'insieme, l'introduzione di una metrica,imponendo la condizione della disuguaglianza triangolare, consente anche la definizione di unarelazione d'ordine tra le distanze dei punti.

Per poter poi definire una relazione più stretta che risponda ad una condizione di classificabilità dei punti, cioè che implichi la possibilità di determinare una "soglia" che definisca una partizione dell'insieme iniziale in due gruppi tale che un elemento si trovi nell'uno o nell'altro gruppo a seconda che la sua distanza da tutti gli altri elementi sia rispettivamente minore o uguale della soglia fissata o maggiore di questa, bisogna ricorrere al concetto di distanza ultrametrica in

base alla quale ciascuna terna di punti definisce un triangolo isoscele con base costituita dal lato più piccolo e con le distanze iniziali modificate secondo determinati criteri.

Obiettivo della cluster analysis e' di identificare un minor numero di gruppi tali che gli elementi appartenenti ad un gruppo siano – in qualche senso – più simili tra loro che non agli elementi appartenenti ad altri gruppi. Il punto di partenza fondamentale è la definizione di una misura di similarità o di distanza tra gli oggetti (cioè tra le righe della matrice dei dati).

L'altro punto fondamentale è la regola in base alla quale si formano i gruppi. Infatti, come già spiegato, a seconda del tipo di dati, si hanno misure diverse. Per dati quantitativi si hanno misure di distanza; per dati qualitativi si hanno misure di associazione.

DATI NUMERICI

Tra i criteri di distanza più frequentemente utilizzati nel caso di oggetti descritti da variabili numeriche ricordiamo:

- La metrica di Minkowsky - Può essere considerata come una metrica generale cui diverse metriche note possono essere ricondotte mediante scelte particolari del valore del parametro. La distanza tra due punti *i* e *h* è definita nel modo seguente:

(1)
$$d_\theta(i,h) = \left(\sum_r |x_{ir} - x_{hr}|^\theta \right)^{1/\theta}$$

in cui la scelta di θ dipende dal rilievo che si vuole dare alle differenze più grandi: maggiore è θ, maggiore è l'enfasi che si dà alle differenze $|x_{ir} - x_{hr}|$ più grandi.

Come casi particolari della metrica di Minkowsky ricordiamo:

- La distanza City block- Detta anche metrica di Manhattan, è così chiamata in quanto rappresenta la distanza che deve coprire un individuo che si muova in una città con strade tra di loro perpendicolari o parallele. Si ottiene dalla formula (1) ponendo θ=1:

$$d_1(i,h) = \sum_{r=1}^{p} |x_{ir} - x_{hr}|$$

- La distanza euclidea - E' probabilmente la più nota e la più utilizzata. Si ottiene come caso particolare della metrica di Minkowsky ponendo θ=2 e nel caso di due variabili rappresenta la distanza di due punti nel piano calcolata facendo ricorso al teorema di Pitagora. In uno spazio a *p* dimensioni la distanza tra i punti *i* e *h* è data da:

(2) $$d_2(i,h) = \sqrt{\sum_{r=1}^{p} (x_{ir} - x_{hr})^2}$$

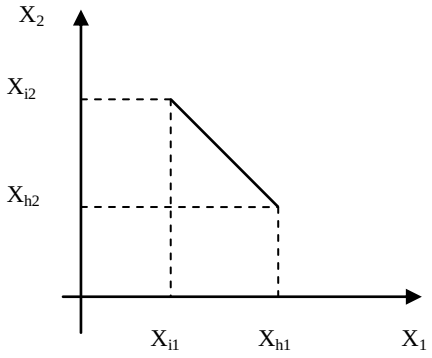


Figura 4 – Distanza tra due punti calcolata con la metrica euclidea.

FREQUENZE

Consideriamo *n* unità statistiche su cui siano state osservate 2 variabili su scala qualsiasi. I dati possono essere raccolti in una tabella a doppia entrata in cui ogni cella riporta la frequenze associata all'incrocio delle corrispondenti modalità delle variabili.

Nelle tabelle così definite la *i*-esima riga (*j*-esima colonna) rappresenta il profilo dell'insieme di unità statistiche che possiede la modalità *i*-esima della prima variabile (*j*-esima della seconda). Il concetto di distanza tra punti risulta diverso da quello definito nelle tabelle di intensità, in cui vengono riportate le determinazioni delle variabili sulle diverse unità. Più precisamente, nel caso di tabelle di frequenze la distanza più utilizzata è la distanza del *chi-quadrato* (o *distanza dei profili*). La distanza tra due righe *i* ed *i'* si può quindi definire come:

$$d^2(i,i') = \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

e la distanza tra due colonne *j* e *j'*:

$$d^2(j,j') = \sum_i \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2$$

avendo indicato con f_{ij}, $f_{i.}$ e $f_{.j}$ rispettivamente le frequenze relative, i marginali di riga ed i marginali di colonna della tabella.

La distanza del chi-quadrato gode della importante proprietà dell’equivalenza distributiva ed è quindi invariante rispetto ai criteri di codifica o al modo di aggregare le unità in gruppi, a condizione che le unità aggregate siano omogenee.

La distanza del chi-quadrato tra due righe (colonne) può essere considerata come una particolare distanza euclidea ponderata con fattori di ponderazione dati dai reciproci dei marginali di colonna (riga)

La metrica del chi-quadrato si distingue quindi dalla tradizionale metrica euclidea poiché tiene conto delle variazioni anche su righe o colonne a bassa numerosità. Poiché con la metrica del chi-quadrato pesa più uno scarto su una modalità poco frequente che uno di pari entità su una ad alta frequenza, tale metrica risulta particolarmente adatta in quelle che mirano ad evidenziare le relazioni locali tra le entità analizzate.

Variabili dicotomiche

Si assuma ora che ciascuna x_{ih} possa assumere valori 0 o 1 per $i = 1,\dots, n$ e $h = 1,\dots, k$. La dissomiglianza tra due osservazioni X_i e X_j può essere rappresentata tramite la seguente tabella

x_j / x_i	1	0
1	a	b
0	c	d

Figura 5 – Schema per il calcolo degli indici di dissimilarità.

in cui

a) numero di variabili che valgono 1 per entrambe le osservazioni;
b) numero di variabili che valgono 1 per la *i*-esima e 0 per la *j*-esima osservazione;
c) numero di variabili che valgono 0 per la *i*-esima e 1 per la *j*-esima osservazione;
d) numero di variabili che valgono 0 per entrambe le osservazioni.

Ovviamente con $a+b+c+d= k$. Questa rappresentazione può essere sintetizzata tramite due indici di dissomiglianza: il *coefficiente di dissomiglianza semplice* e il *coefficiente di Jaccard*.

Il coefficiente di dissomiglianza semplice. E’ dato dalla proporzione delle variabili che risultano discordanti:

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

Il coefficiente di Jaccard risulta indicato per variabili dicotomiche asimmetriche, che indicano la presenza di una data caratteristica. In tal caso l’assenza della caratteristica da entrambe le unità non dovrebbe contribuire ad aumentarne il grado di somiglianza:

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

Variabili miste

Se vengono rilevate variabili di natura diversa (qualitative, quantitative, binarie) la perdita di informazioni che implicherebbe la riduzione di tutte le variabili alla scala di precisione inferiore può essere evitata applicando l’indice di Gower (1971):

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

dove se la *h*-esima variabile è *quantitativa* ed *R*(*h*) è il suo campo di variazione, si ha

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

mentre se `e *qualitativa*

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

ed in generale

$$\delta_{ijh} = \begin{cases} 1 & \text{se si conoscono i valori dell'}h\text{-esima variabile per le osservazioni } i\text{-esima e } j\text{-esima (serve quando vi sono dati mancanti);} \\ \mathbf{0} & \text{nel caso contrario, oppure in caso di accordo 0/1 per variabili binarie di tipo presenza/assenza.} \end{cases}$$

6. TECNICHE DI ANALISI DEI GRUPPI

Tecniche gerarchiche di analisi dei gruppi

Con questo nome si fa riferimento ai criteri per la creazione di partizioni annidate dell'insieme di osservazioni di partenza. Tali criteri permettono di esplorare la struttura di raggruppamento con riferimento a livelli variabili di omogeneità all'interno dei gruppi. La considerazione delle sole partizioni annidate, piuttosto che di tutte le partizioni possibili, riduce considerevolmente i tempi dell'analisi. D'altro canto un errore commesso nella fase iniziale della classificazione non può più essere messo in discussione.

Un elemento cruciale per la comprensione del funzionamento delle tecniche gerarchiche è la rappresentazione grafica della struttura di raggruppamento tramite *diagrammi ad albero* o *dendrogrammi* (Fig. 6). Sezionando il dendrogramma in corrispondenza di un certo livello di dissomiglianza si ottiene una partizione in gruppi disgiunti e omogenei dell'insieme di unità.

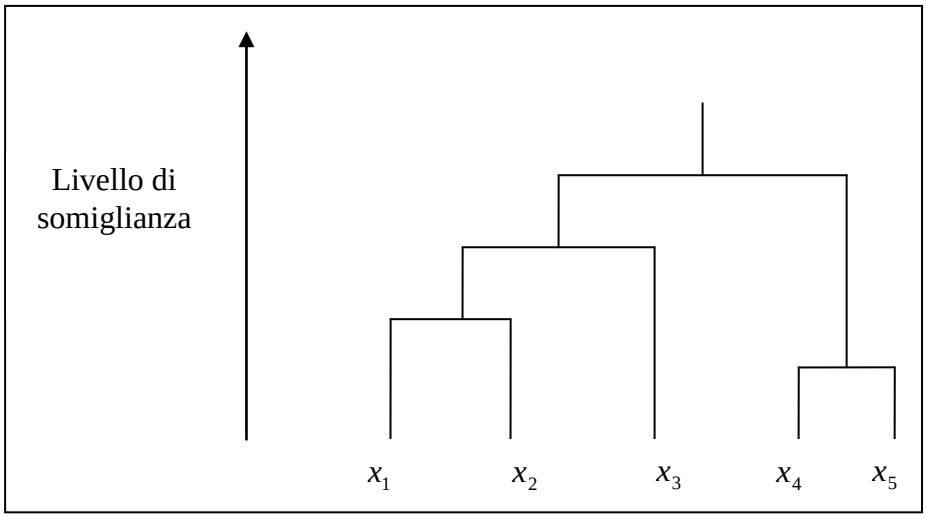


Figura 6 - Esempio di dendrogramma

Vengono definiti *algoritmi aggregativi* quelli che procedono per aggregazioni successive di unità, ovvero dalle foglie alla radice del diagramma ad albero. Viceversa si dicono *algoritmi scissori* quelli che suddividono il collettivo di partenza in gruppi di dimensioni via via più ridotte. Di conseguenza, nel caso in cui si voglia individuare un numero ridotto di gruppi, gli algoritmi scissori risultano più efficienti in termini di numero di passi da compiere.
Nonostante la complessità di calcolo sia in genere di gran lunga superiore negli algoritmi scissori che in quelli aggregativi, i primi vengono comunque preferiti ai secondi.

Tecniche gerarchiche aggregative.

Dal collettivo non suddiviso in gruppi si procede per aggregazioni successive generando gruppi sempre più numerosi. Il procedimento di raggruppamento parte dalla matrice di dissomiglianza tra elementi e procede iterativamente in due passi:

(i) raggruppando gli elementi più somiglianti;

(ii) calcolando la matrice di dissomiglianza tra gruppi e/o elementi, avendo fissato un criterio per stabilire la distanza dei gruppi dai singoli elementi e/o dagli altri gruppi.

Il procedimento si arresta quando tutti gli elementi sono aggregati in un unico cluster.
Una volta stabilito l'indice di dissomiglianza o la distanza tra le osservazioni, ciò che differenzia le diverse tecniche gerarchiche di clustering è esclusivamente il criterio per stabilire la dissomiglianza o la distanza tra i diversi cluster (in quanto segue, per brevità, si parla di distanze sia nel caso di osservazioni quantitative che qualitative).

Metodo del legame semplice o del vicino più prossimo. La distanza tra due gruppi è data dalla *minore* delle distanze tra gli elementi. Un possibile effetto collaterale di questo metodo è il concatenamento tra unità appartenenti a gruppi diversi. Questo metodo è definito da Benzècri criterio del salto minimo, con tale legame un oggetto verrà unito al gruppo per il quale risulta minima la distanza dall’elemento più vicino. Quindi la distanza di c_1 da c_2 è data dalla più piccola distanza tra un elemento di c_1 e un elemento di c_2.

$d(c_1, c_2) = \min\{d(e_i, e_j)\} : e_i \in c_1; e_j \in c_2$

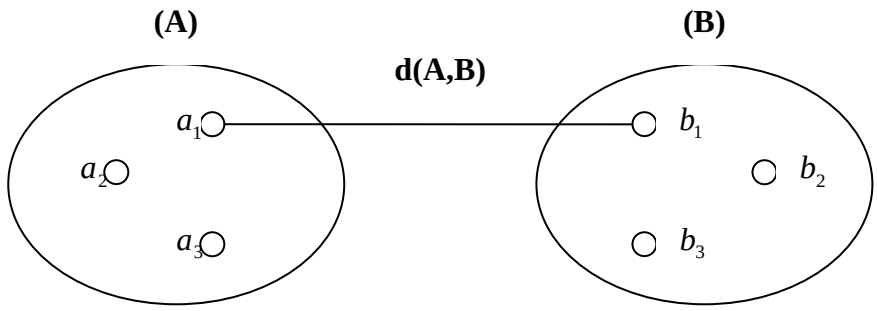


Figura 7 – Distanza tra due gruppi con il metodo del legame semplice

Metodo del legame completo. La distanza tra due gruppi `e data dalla *maggiore* delle distanze tra gli elementi, ossia dal diametro della più piccola sfera che include il gruppo ottenuto aggregando i due gruppi (metodo che dà luogo a gruppi di forma sferica). Con questo metodo un oggetto verrà unito al gruppo per il quale la distanza massima da un elemento risulta minima. I gruppi definiti con questo criterio sono in genere numerosi ma molto omogenei, dando questo metodo più peso alla proprietà della coesione interna che non a quella dell’isolamento dei gruppi. Quindi la distanza di c_1 da c_2 è data dalla distanza massima tra un elemento di c_1 e un elemento di c_2.

$d\,(c_1, c_2) = \max\{d\,(e_i, e_j)\} : e_i \in c_1; e_j \in c_2$

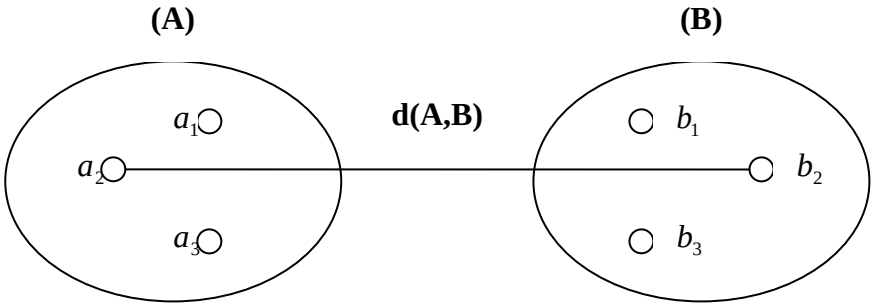
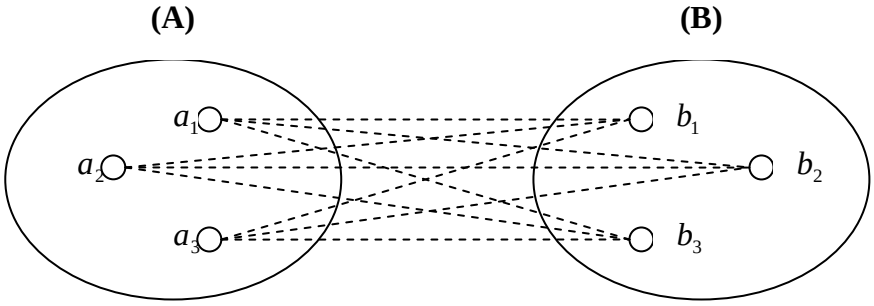


Figura 8 – Distanza tra due gruppi con il metodo del legame completo

<u>*Metodo del legame medio o della media di gruppo*</u>. : questo metodo fu introdotto alla fine degli anni ’50 come metodo intermedio tra quello del legame semplice e del legame completo, assegna un oggetto al gruppo per il quale risulta minima la distanza media, intendendo con questa espressione la media delle distanze da tutti i punti del gruppo. Il metodo del legame medio è quello che genera la minore distorsione delle distanze iniziali. Quindi la distanza di c_1 da c_2 è data dalla media delle distanze di tutti i punti di c_1 da tutti i punti di c_2.



$$d(A,B) = \frac{1}{n_A . n_B} \sum_i \sum_j d(a_i, b_j)$$

Figura 9 – Distanza tra due gruppi con il metodo del legame medio

<u>*Metodo del centroide*</u>. Il centroide di ciascun gruppo è definito come il punto che ha per coordinate la media delle coordinate degli elementi del gruppo. La distanza tra due gruppi è data dalla distanza

euclidea tra i due centroidi corrispondenti. Ad ogni passo della procedura vengono aggregati i gruppi per i quali la distanza euclidea tra i centroidi risulta minima. Questo metodo parte invece dalla considerazione che ogni oggetto può essere visto come in uno spazio euclideo. Ad ogni passo vengono uniti i gruppi i cui centroidi risultano più vicini. La distanza di c_1 da c_2 è data dalla distanza tra i baricentri di c_1 e c_2.

$d(c_1, c_2) = d(\bar{e}_1, \bar{e}_2)$ $\quad e_i \Rightarrow$ *baricentro* di c_i (i = 1,2).

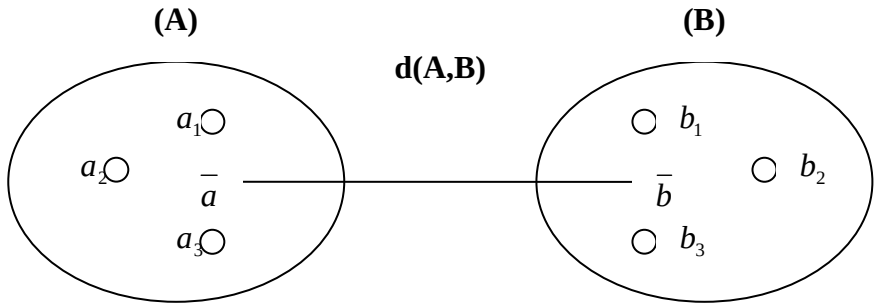


Figura 10 – Distanza tra due gruppi con il metodo dei centroidi

<u>*Metodo di Ward*</u>. questo metodo è basato sulla minimizzazione della varianza all'interno dei gruppi. Ad ogni passo vengono calcolate le devianze associate a tutti i raggruppamenti possibili e viene effettuata l'aggregazione che dà luogo al gruppo avente devianza minima. La distanza tra due gruppi è data dalla differenza tra la devianza complessiva e la somma delle devianze interne ai due gruppi, ovvero dall'incremento della devianza entro i gruppi dovuto all'aggregazione in questione.
Per meglio chiarire i principi del metodo, sappiamo che l'inerzia totale della nube degli *n* punti è data dalla somma dei quadrati delle distanze dal baricentro ***g,*** ciascuna distanza ponderata dalla massa m_i di ciascun punto:

$$\Im = \sum_{i=1}^{n} m_i \cdot d^2(x_i, g)$$

con:

$$m = \sum_{i=1}^{n} m_i \;\; ; \quad g = \frac{1}{m}\sum_{i=1}^{n} m_i x_i \;\; ; \qquad d^2(x_i, g) = \|x_i - g\|^2 \;\; .$$

A questo punto il teorema di Huyghens ci consente di decomporre l’inerzia totale in una parte relativa alla variabilità *entro* i gruppi ed una parte relativa alla variabilità *tra* i gruppi:

$$\Im = \sum_{j=1}^{k} m_j \left\| g_j - g \right\|^2 + \sum_{j=1}^{k} \sum_{i=1}^{n_k} m_i \left\| x_i - g_j \right\|^2 \ .$$

Poiché l’inerzia totale è costante, obiettivo della partizione è quello di minimizzare la quota di variabilità interna ai gruppi, massimizzando al contempo la variabilità tra i gruppi, così da ottenere classi omogenee al loro interno e ben separate l’una dall’altra.

Inerzia totale = inerzia entro le classi – inerzia tra le classi

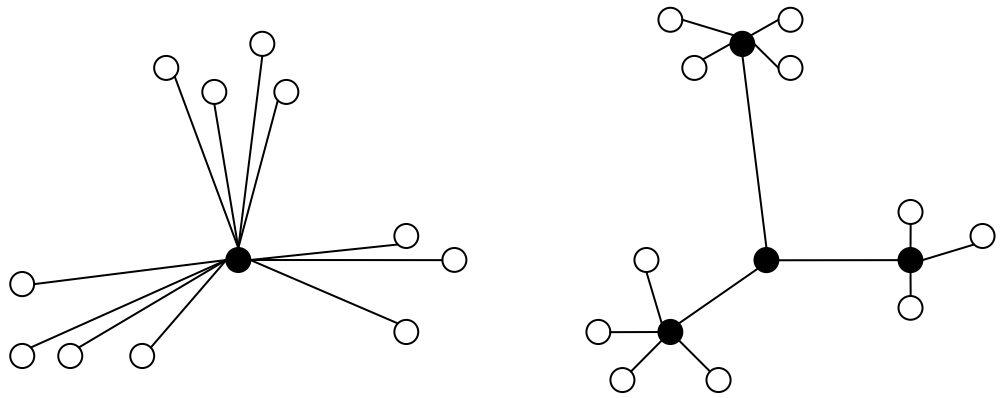


Figura 11 – Decomposizione dell’inerzia secondo la relazione di Huyghens.

Al primo passo, si considerano n classi, ciascuna formata da un individuo mentre al passo conclusivo tutti gli individui sono raggruppati in un’unica classe: sono queste le situazioni limite in cui la varianza tra i gruppi è rispettivamente massima e nulla. L’algoritmo di Ward aggrega, ad ogni passo intermedio, gli oggetti (gruppi o unità) che determinano la perdita di inerzia minima. Più precisamente, ad ogni passo s (s=2,…,n) l’inerzia tra le classi diminuisce della quantità ΔI (e l’inerzia entro le classi aumenta della stessa quantità); queste quantità possono essere considerate come degli indici di dissimilarità che verificano la relazione:

$$\sum_{s=2}^{n} \Delta_s = I \ .$$

Tecniche gerarchiche scissorie.

Considerano tutte le unità aggregate in un unico cluster e ne cercano la suddivisione in due gruppi che massimizza una certa funzione obiettivo. Ad ogni passo successivo viene effettuata la migliore suddivisione dei gruppi già individuati. Il procedimento ha termine quando ciascun cluster risulta formato da una sola osservazione (nel caso di attributi dicotomici le tecniche gerarchiche scissorie prendono anche il nome di *tecniche di segmentazione binaria*).

<u>*Metodo K-means*</u>. Il metodo *k*-means (Mc Queen, 1967) inizia considerando la bipartizione del collettivo che minimizza la devianza interna ai gruppi e procede, ad ogni passo, suddividendo il gruppo avente devianza maggiore, in modo che la devianza interna complessiva risulti minima (il che equivale a massimizzare la distanza tra i centroidi dei gruppi). Tale impostazione è applicabile anche qualora si facciano $K > 2$ suddivisioni all'interno di ogni gruppo. Questo criterio viene utilizzato anche nell'ambito delle tecniche non gerarchiche di raggruppamento.

<u>*Metodo di Edwards e Cavalli Sforza*</u> *(1965)*. Dette $W = W_1 + \ldots + W_r$ e $B = T\text{-}W$ rispettivamente la matrice delle devianze e codevianze all'interno di e tra *r* gruppi, la bipartizione viene iterata in modo da massimizzare la traccia di *B*. Questo metodo risulta equivalente al metodo *k*-means.

Tecniche non gerarchiche di analisi dei gruppi

Le tecniche non gerarchiche di analisi dei gruppi consistono in generale nei seguenti due passi:

(i) si individua una *partizione provvisoria* del collettivo in un certo numero di gruppi (tramite tecniche gerarchiche di analisi dei gruppi o informazioni a priori, specificando o meno il numero di gruppi della partizione);

(ii) si ottimizza una *funzione obiettivo* modificando l'assegnazione (tramite diversi metodi di programmazione matematica).

Le diverse tecniche non gerarchiche si differenziano per le caratteristiche delle funzioni obiettivo. Le funzioni obbiettivo considerate nel seguito sono tutte invarianti per trasformazioni lineari dei dati.

La *minimizzazione della traccia della matrice W* delle devianze interne ai gruppi coincide con i criteri di Edwards e Cavalli Sforza e *K*-means già trattati nell'ambito delle tecniche gerarchiche scissorie.

In alternativa si può considerare la *massimizzazione della traccia di Hotelling*, ovvero della traccia della matrice BW^{-1}, che sta per la matrice di varianza/covarianza tra i gruppi diviso la matrice di varianza/covarianza interna ai gruppi.

Con il *metodo di Forgy*(1965) le unità più vicine ai centroidi dei gruppi vengono incluse negli stessi gruppi:

- *(i)* si calcolano i centroidi dei gruppi della partizione iniziale;
- *(ii)* si confrontano tutte le unità con i centroidi calcolati;
- *(iii)* si ricollocano le unità nei gruppi il cui centroide risulta più vicino;
- *(iv)* si ricalcolano i centroidi;
- *(v)* si riprende da (*iii*).

Questo metodo è sostanzialmente analogo alla minimizzazione della traccia di *W*.

Qualora il numero dei gruppi *r* che formano la partizione non sia prestabilito a priori, occorre prefissare un parametro sostitutivo, come ad esempio una soglia di distanza sufficiente per aggregare le osservazioni, ovvero la numerosità dei gruppi.

7. SCELTA TRA LE TECNICHE DI ANALISI DEI GRUPPI

Si possono individuare alcuni requisiti di carattere generale che devono essere posseduti da una buona tecnica di raggruppamento e che riguardano la sua *oggettività*, in quanto deve garantire risultati identici se applicata sugli stessi dati da persone diverse indipendentemente, la *stabilità* dei risultati a piccole variazioni nei dati (ad esempio l'eliminazione di un'unità dall'analisi), l'*informazione* del raggruppamento finale e la *semplicità* dal punto di vista algoritmico, con riferimento alla rapidità dell'esecuzione dei calcoli ed alla quantità di memoria necessaria.

Bisogna comunque sottolineare come in generale i risultati di una strategia di raggruppamento dipendano oltreché dalla tecnica utilizzata, anche dal tipo di distanza e dall’uso di osservazioni grezze o standardizzate.

La scelta tra l’utilizzo delle tecniche gerarchiche o di quelle non gerarchiche deve tenere conto necessariamente degli inconvenienti causati dalla rigidità delle strutture di raggruppamento annidate: un’aggregazione impropria effettuata nei primi passi dell’analisi viene portata avanti sino alla fine. Per questo le tecniche gerarchiche, oltre a risentire molto della presenza di errori di misura e/o osservazioni anomale, portano in generale a gruppi meno omogenei di quelli ottenibili attraverso le tecniche non gerarchiche.

D’altro canto dal punto di vista del calcolo le prime risultano molto meno dispendiose delle seconde (in caso di r gruppi $\binom{r}{2}$ contro $2^{r-1}-1$confronti per le tecniche non gerarchiche).

Nel caso di un numero molto elevato di unità da raggruppare va anche considerata la quantità di memoria necessaria per effettuare l’analisi: mentre per le tecniche gerarchiche bisogna tenere in memoria almeno $n(n - 1)/2$ numeri (tanti quanti sono gli elementi della matrice di dissomiglianza delle singole osservazioni), le tecniche non gerarchiche richiedono di tenere in memoria la sola matrice dati.

Dalle considerazioni applicative sin qui fatte le tecniche non gerarchiche sembrerebbero più convenienti. In realtà esse risultano troppo dispendiose dal punto di vista computazionale per le analisi di collettivi molto numerosi, essendo la convergenza ad una soluzione ottimale spesso molto lenta. Inoltre come per tutti i metodi basati sui quadrati delle osservazioni, l’eventuale presenza di dati anomali può costituire un serio problema.

Infine i risultati delle tecniche non gerarchiche sono modesti se non si dispone di una buona partizione iniziale fornita, ad esempio, da un’analisi di raggruppamento gerarchica preliminare.

Nell’ambito delle tecniche gerarchiche aggregative la scelta deve essere effettuata nel rispetto della natura dei dati da analizzare. Infatti il metodo del legame singolo causa il concatenamento tra le unità. Esso è pertanto adatto ad includere dati di dimensioni anomale nei gruppi, oppure per gruppi che si sospetta abbiano forma allungata. Il metodo del legame completo risulta invece particolarmente raccomandato per variabili qualitative, mentre il metodo di Ward tende ad isolare le unità che presentano valori estremi anche di una sola variabile.

Inoltre la bontà di diverse soluzioni gerarchiche può essere valutata comparativamente in termini della distorsione indotta dalla considerazione dei raggruppamenti in luogo delle singole

osservazioni, tramite il *coefficiente dicorrelazione cofenetico* (Sokal e Rohlf, 1962). Siano detti cofenetici ed indicati con d_{ij}^{*} i valori di prossimità deducibili dal procedimento gerarchico di classificazione, ovvero dal dendrogramma (la prossimità tra due osservazioni è data dal livello di dissomiglianza in corrispondenza del quale entrano a far parte dello stesso gruppo).

$$R_c = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(d_{ij}-\overline{d}\right)\left(d_{ij}^{*}-\overline{d}^{*}\right)}{\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(d_{ij}-\overline{d}\right)^2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(d^{*}{}_{ij}-\overline{d}^{*}\right)^2\right]^{\frac{1}{2}}}$$

Nell’espressione precedente $\overline{d}$ e $\overline{d}^{*}$ rappresentano rispettivamente la media delle dissomiglianze semplici e cofenetiche tra le *n* osservazioni. L’indice R_c diminuisce al crescere della distorsione del dendrogramma e i suoi valori sono in genere compresi tra 0,6 e 0,95.

8 DETERMINAZIONE DEL NUMERO DI GRUPPI

Poiché raramente nell’analisi dei gruppi si conosce il numero dei raggruppamenti da individuare, esistono dei criteri per la determinazione dello stesso.

Verifica della significatività del raggruppamento (Beale,1969). Viene verificata l’ipotesi che passando da *r*- 1 ad *r* gruppi si abbia una riduzione significativa della devianza interna ai gruppi (ovvero che i centroidi degli *r* gruppi siano significativamente distanti). Indicando con D_r^2 la traccia (somma degli elementi sulla diagonale) della matrice delle devianze e codevianze interne *W* calcolata in corrispondenza di una partizione in *r* gruppi, tale ipotesi viene verificata tramite la statistica

$$F = \left(\frac{D_{r-1}^2 - D_r^2}{D_r^2}\right)\left[\frac{n-r-1}{n-r}\left(\frac{r-1}{r}\right)^{\frac{k}{2}} - 1\right]^{-1}$$

la cui regione critica asintotica è ottenuta considerando la coda destra della distribuzione F di Fisher con k e $k(n - r)$ gradi di libertà.

Verifica della significatività degli autovalori. Per le tecniche non gerarchiche basate sulla traccia delle matrici W e BW^{-1} si può costruire il test del rapporto di verosimiglianze generalizzato per la verifica della significatività degli autovalori delle stesse matrici.
L’uso di test per la verifica di ipotesi in questo contesto è piuttosto dibattuto. Infatti, al crescere dell’ampiezza n del collettivo, i centroidi dei gruppi risultanti da qualsiasi procedimento classificatorio risultano quasi sempre significativamente diversi. Un uso preferibile consiste nella verifica della significatività di una serie di soluzioni corrispondenti a diversi valori di r in modo da individuare quella maggiormente significativa. In alternativa si può fare ricorso a verifiche empiriche basate su indici sintetici di derivazione euristica o su tecniche di simulazione.
Nel caso dell’uso di tecniche gerarchiche un modo per determinare il numero di gruppi da considerare consiste nel sezionare il dendrogramma in corrispondenza del massimo scarto tra i livelli di prossimità ai quali sono avvenute le aggregazioni.

9. SCELTA DEL METODO E QUALITÀ DEI RISULTATI

I metodi di classificazione automatica, ed in particolare i metodi gerarchici aggregativi, hanno trovato ampia applicazione in molti campi grazie anche alla possibilità di essere utilizzati su collettivi molto numerosi. Diversi autori hanno affrontato il problema della scelta del metodo ottimale, chi utilizzando un approccio più teorico chi affidandosi ad argomentazioni prevalentemente empiriche, anche se la validità dei raggruppamenti risulta di volta in volta qualificata più in base a considerazioni di tipo soggettivo e a criteri di migliore interpretabilità del fenomeno analizzato che non a rigidi schemi teorici di riferimento.
Volendo definire una regola molto generale, si può dire che la scelta del metodo non dovrebbe compromettere la validità della classificazione in termini di stabilità dei risultati, vale a dire che, quale che sia la scelta effettuata, il metodo non dovrebbe risentire particolarmente di:

- lievi errori nella registrazione dei dati;
- aumento limitato del numero degli oggetti;
- aumento limitato del numero delle variabili.

Inoltre, ciascun metodo di classificazione considera, implicitamente, delle ipotesi sulla “forma” o sulla “dimensione” dei gruppi. Ad esempio, il metodo del legame semplice può fornire gruppi di forma e dimensione qualsiasi se e soltanto se i gruppi stessi sono isolati e non uniti da “catene” di oggetti intermedi, nel qual caso, invece, il metodo tenderà a produrre gruppi allungati alle estremità; al contrario, il metodo di Ward tende a formare gruppi di uguale dimensione e di forma ipersferica.
Ogni criterio di classificazione mira, quindi, ad individuare particolari tipi di gruppi e, se la struttura dei dati fosse nota, si dovrebbe scegliere quel metodo o criterio che meglio la evidenzia. Normalmente, però, tale struttura non è nota, anzi si può dire che l’analisi viene effettuata proprio per individuarla ed è quindi presente il rischio di utilizzare un metodo poco appropriato. È allora consigliabile ripetere la classificazione utilizzando metodi diversi e confrontare i risultati.
Volendo schematizzare le principali differenze tra le procedure gerarchiche e non gerarchiche si può dire che:

- i metodi non gerarchici forniscono risultati la cui interpretazione è generalmente più semplice perché costituiti da un’unica partizione. Nel caso dell’analisi gerarchica, invece, la lettura dei risultati si presenta più complessa ma anche più ricca di possibili approfondimenti. Un’analisi gerarchica fornisce infatti una strutturazione dei dati più analitica che richiede una maggiore attenzione in sede di interpretazione. In particolare, la lettura della gerarchia ottenuta dovrà essere fatta sia in senso “verticale” che “orizzontale”: la prima fornisce un’informazione sul *come* si raggruppano gli elementi, la seconda definisce, ad un determinato livello della gerarchia, *quali* elementi fanno parte dei diversi gruppi. Ciò consente un’interpretazione globale che include anche i diversi passi attraverso i quali si è giunti alla partizione scelta. D’altra parte, una volta scelto il livello di aggregazione, un metodo gerarchico viene di fatto interpretato come un metodo non gerarchico.

- Le tecniche gerarchiche possono risultare meno efficaci di quelle non gerarchiche, in particolare di quelle che utilizzano criteri di aggregazione che generano partizioni caratterizzate da gruppi con forte omogeneità interna.

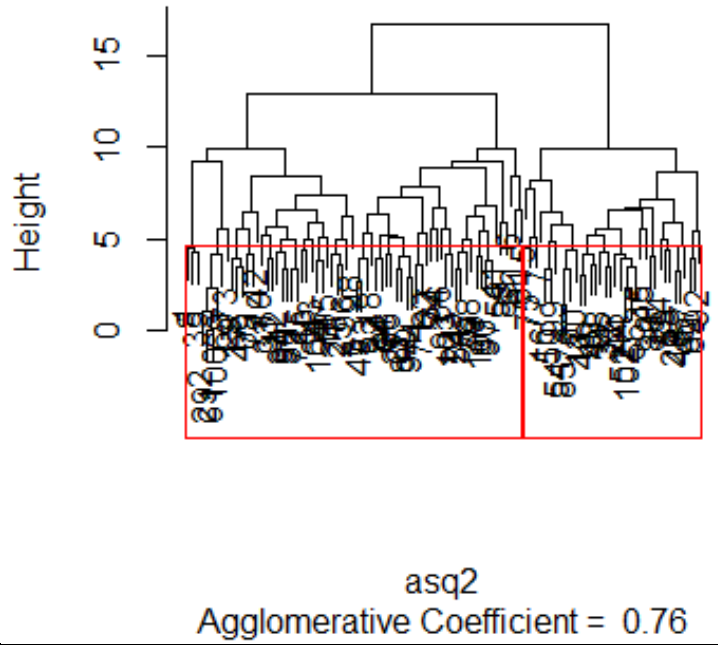
- Quando si hanno molte unità il costo computazionale della tecniche gerarchiche aumenta. Le $\left(\frac{n(n-1)}{2}\right)$ distanze che è necessario calcolare per una tecnica gerarchica ascendente richiedono certamente meno risorse, in termini di tempo di calcolo e di memoria richiesta, delle $n \times k$ necessarie per una tecnica non gerarchica con *k* gruppi.

- La soluzione di una tecnica gerarchica può essere influenzata dalle aggregazioni effettuate ai primi stadi. In particolare, la logica delle partizioni nidificate comporta che un’aggregazione impropria effettuata nei primi stadi possa produrre delle distorsioni anche nelle successive aggregazioni. D’altra parte, una tecnica non gerarchica può non avere molto senso se non si hanno delle informazioni a priori che possano indirizzare nella scelta del numero di gruppi. Una possibile soluzione è allora quella di eseguire una procedura non gerarchica dopo aver analizzato i risultati di una gerarchica. La valutazione dei risultati di un metodo di classificazione gerarchico può inoltre essere effettuata confrontando la matrice delle ultrametriche legata al particolare criterio scelto con quella delle distanze calcolate con la metrica iniziale.

10. APPLICAZIONE DI UN ALGORITMO DI CLASSIFICAZIONE GERARCHICA SULLA SODDISFAZIONE DEI VIAGGIATORI IN TRENO

Algoritmo in ambiente di programmazione R chiamato “agnes” con library (cluster). Il criterio di aggregazione usato è quello completo. Dati: 105 viaggiatori su ferro, su tratte di viaggio campane.
L’indice di agglomerazione è abbastanza elevato.

Istogramma e taglio sull’albero



Caratteristiche dei due gruppi rispetto alle 28 variabili che hanno caratterizzato il servizio di trasporto ferroviario.

>	aggregate(asq2,list(cutree(asq2agnes,2)),median)					
	Group.1	transport	parking	vfm_parking	baggage_carts	airline_wait
1	1	3.0	3	3	3	4
2	2	2.5	2	2	2	3
	trainline_effic	trainline_court	security_staff	through_inspect	waiting_inspect	
1	4	4	4	4	4	
2	3	3	3	3	3	
	feel_safe	finding_way	screens	walking	airport_staff	restaurants
1	4	3	5	4	4	4
2	3	2	3	3	3	3
	vfm_restaurant	bank	shopping	vfm_shopping	washrooms_av	washrooms_clean
1	3	4	4.0	3	3.5	4
2	2	2	2.5	2	3.0	2
	waiting_areas	cleanliness	ambience	overall_sat	gender	age
1	4	3	3	3	1	3
2	3	2	2	2	1	4

Il cluster 1, composto da 68 viaggiatori di età minore di 34 anni, è quello più soddisfatto delle diverse 28 caratteristiche di qualità del servizio, differentemente dal cluster 2 (meno soddisfatti) composto da 36 viaggiatori con età maggiore di 34 anni.

Grafico Silhouette per la valutazione dell’omogenità dei due cluster