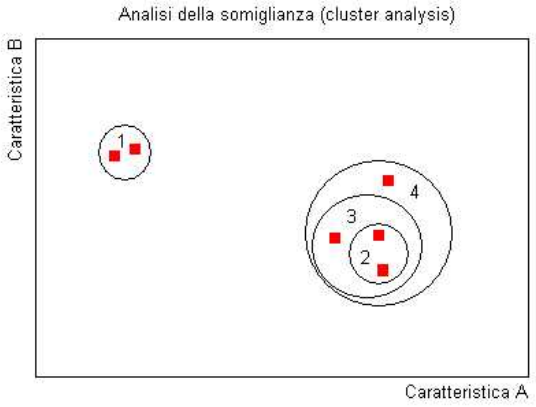
18. Analisi della somiglianza (cluster analysis)

I problemi di classificazione rivestono un ruolo centrale nella scienza. E classificare gli oggetti in base a criteri oggettivi, basati su quantità misurabili, è lo scopo fondamentale della cluster analysis.

Il concetto di base è semplice: raggruppare oggetti omogenei (organismi biologici, minerali, eccetera) in insiemi (cluster), partendo dagli oggetti più simili, aggiungendo progressivamente gli oggetti più dissimili. All'inizio del processo di classificazione ad ogni oggetto corrisponde un cluster (e viceversa). In questo stadio tutti gli oggetti sono considerati dissimili tra di loro.



Al passaggio successivo i due oggetti più simili sono raggruppati in un unico cluster. Il numero dei cluster risulta quindi pari al numero di oggetti diminuito di 1. Il procedimento viene ripetuto ciclicamente, fino ad ottenere (all'ultimo passaggio) un unico cluster. Nella figura viene presentato un esempio di clusterizzazione, I due oggetti più vicini (gli oggetti sono rappresentati in un diagramma cartesiano in base a due caratteristiche, A e B) sono raggruppati nel cluster 1. Successivamente si forma un secondo cluster (cluster 2) che raggruppa gli oggetti più vicini tra loro dopo i primi due. Un ulteriore oggetto viene raggruppato nel cluster tre, che viene in tal modo a comprendere, oltre a quest'ultimo oggetto, i due oggetti precedentemente raggruppati nel cluster 2. Infine i tre oggetti precedentemente raggruppati nel cluster 3 confluiscono nel cluste 4 insieme all'oggetto più vicini. Stabilire a quale livello di aggregazione degli oggetti fermarsi, e quindi quali conclusioni trarre, dipende in larga parte dal giudizio di merito dell'utilizzatore. Per questo la cluster analysis riveste un ruolo centrale, in statistica, limitatamente alla analisi esplorativa dei dati (nel caso specifico sembrerebbe che in definitive gli oggetti finiscano con il confluire in due diverse "famiglie", significativamente diverse tra di loro).

Si consideri il seguente esempio, relativo a 10 calcoli delle vie urinarie, analizzati per la loro composizione in calcio, fosfato, ossalato e magnesio (dati simulati e valori espressi in unità di misura arbitrarie).

CALCOLO	CALCIO	FOSFATO	OSSALATO	MAGNESIO
1	99	81	69	61
2	78	65	53	43
3	81	66	38	54
4	45	23	19	16
5	44	18	24	19
6	102	83	72	66
7	83	68	49	45
8	74	71	41	57
9	38	19	22	14
10	48	14	21	12

All'inizio del processo di classificazione (clustering) ad ogni oggetto corrisponde un cluster (e viceversa). In questo stadio tutti gli oggetti sono considerati dissimili (diversi) tra di loro. Al passaggio successivo i due oggetti più simili sono raggruppati in un unico cluster. Il numero dei

cluster risulta quindi pari al numero di oggetti - 1. Il procedimento viene ripetuto ciclicamente, fino ad ottenere (all'ultimo passaggio) un unico cluster.

Stabilire a quale livello di aggregazione degli oggetti fermarsi, e quindi quali conclusioni trarre, dipende esclusivamente dai giudizio di merito dell'utilizzatore. Per questo la cluster analysis riveste un ruolo centrale, in statistica, limitatamente alla analisi esplorativa dei dati (in effetti il livello a cui fermarsi trarre le conclusioni a questo punto non è più quantitativo, e quindi non è più oggettivo).

La prima cosa che si fa nell'analisi della somiglianza è quella di calcolare le distanze euclidee. La distanza può essere calcolata in vari modi: quello qui seguito prevede di calcolare la distanza tra tutte le possibili coppie di punti. Nel caso di due variabili (come per esempio sarebbe stato se si fossero avute, per ciascun calcolo, le sole misure di calcio e fosfato) mediante il teorema di Pitagora. Che peraltro può essere esteso dal piano cartesiano, bidimensionale, ad uno spazio tridimensionale (nel caso di tre misure sullo stesso campione) e a uno spazio n-dimensionale (nel caso di n misure sullo stesso campione).

Nel caso dell'esempio illustrato la matrice delle distanze è la seguente:

Caso	1	2	3	4	5	6	7	8	9	10
1	0,00	4,72	5,56	13,50	13,21	0,94	4,51	5,44	14,05	13,87
2	4,72	0,00	2,79	8,85	8,60	5,63	0,98	2,81	9,39	9,28
3	5,56	2,79	0,00	8,85	8,74	6,32	2,12	1,20	9,58	9,44
4	13,50	8,85	8,85	0,00	1,03	14,38	9,12	9,14	1,12	1,21
5	13,21	8,60	8,74	1,03	0,00	14,07	8,95	8,99	1,08	1,27
6	0,94	5,63	6,32	14,38	14,07	0,00	5,41	6,17	14,92	14,75
7	4,51	0,98	2,12	9,12	8,95	5,41	0,00	2,39	9,75	9,58
8	5,44	2,81	1,20	9,14	8,99	6,17	2,39	0,00	9,79	9,81
9	14,05	9,39	9,58	1,12	1,08	14,92	9,75	9,79	0,00	1,41
10	13,87	9,28	9,44	1,21	1,27	14,75	9,58	9,81	1,41	0,00

La matrice delle distanze è formata da un numero di righe e di colonne uguale, e pari al numero di casi in esame (i casi sono numerati in ordine progressivo: il caso 1 corrisponde alla prima riga di dati, il caso 2 alla seconda, e così via). La matrice contiene le distanze tra tutte le possibili coppie di casi. Notare come essa sia simmetrica rispetto alla diagonale. Le distanze sono espresse come deviata normale standardizzata (*z*). Si noti che la matrice è simmetrica, e che la diagonale assume valori uguali a zero (la distanza euclidea tra un calcolo e sé stesso è ovviamente nulla).

La corrispondente matrice dei cluster appare così

Caso	Cluster								
10	0	0	0	0	0	6	6	6	9
5	0	0	3	4	4	6	6	6	9
4	0	0	3	4	4	6	6	6	9
9	0	0	0	4	4	6	6	6	9
6	1	1	1	1	1	1	1	8	9
1	1	1	1	1	1	1	1	8	9
2	0	2	2	2	2	2	7	8	9
7	0	2	2	2	2	2	7	8	9
8	0	0	0	0	5	5	7	8	9
3	0	0	0	0	5	5	7	8	9

0.94 0.98 1.03 1.08 1.20 1.21 2.12 4.51 8.60
Distanza euclidea

La matrice dei cluster contiene, per ciascuno dei casi in esame (righe) il/i cluster nei quali il caso confluisce. I cluster sono numerati in ordine progressivo di formazione, da sinistra verso destra. Ogni colonna rappresenta in questo modo un livello crescente di aggregazione dei casi in base alla loro somiglianza. Per ogni colonna è riportata la distanza (standardizzata) alla quale si è formato l'ultimo cluster.
Il primo cluster, comprende gli oggetti numero 6 e numero 1, che hanno una distanza euclidea di 0.94 (controllare anche la matrice delle distanze). Quindi gli oggetti 1 e 6 sono i più simili tra loro in assoluto. Ma anche gli oggetti2 e 7 sono molto simili tra di loro, avendo una distanza euclidea di 0.98. Notare come all'ottavo passaggio si siano formati due cluster, il primo comprendente gli oggetti 10, 5 4 e 9, il secondo comprendente gli oggetti 6, 1, 2, 7, 8 e 3: perché questi due cluster confluiscano si arriva ad una distanza euclidea di 8.60. Quindi questi due gruppi di oggetti sembrano rappresentare due famiglie ben distinte per composizione.