Numero 7 - 1994

RATIO MATHEMATICA

Rivista di Matematica Applicata all'Economia e all'Ingegneria

a cura di

Franco Eugeni e Antonio Maturo

Numero 7 - 1994

RATIO MATHEMATICA

Rivista di Matematica Applicata all'Economia e all'Ingegneria

a cura di

Franco Eugeni e Antonio Maturo

INDICE

PREFAZIONE

Il presente volume contiene i testi di alcune selezionate conferenze e comunicazioni tenute in convegni dedicati alla ricerca in Didattica della Matematica quali il 1° Convegno Interregionale su Probabilità, Statistica e Crittografia nell'insegnamento nelle scuole medie superiori, tenutasi ad Isernia nel 1992, i Convegni Nazionali Annuali della Mathesis, Società di Scienze Fisiche e Matematiche del 1992 e del 1993, per i quali, non sono stati pubblicati gli Atti ed il Congresso Internazionale ICME - 7 tenutosi a Québec nel 1992. Inoltre contiene alcuni articoli inviati alla rivista e ritenuti degni di pubblicazioni dal comitato scientifico.

Abbiamo ritenuto importante trovare una maniera per rendere noti alcuni degli interessanti contributi dati dagli insegnanti di Matematica in questi convegni e a tale scopo abbiamo trovato un sostegno da parte dell'Amministrazione dell'Istituto Mecenate di Pescara.

È nostra convinzione, condivisa dai nostri colleghi dell'Università che un Istituto Superiore parificato si possa qualificare ad un livello superiore a varie scuole statali se impegna una parte delle sue energie alla Ricerca Didattica e riesce, in tale settore, a stabilire una collaborazione permanente con l'Università.

L'intera organizzazione dell'Istituto Mecenate, ha mostrato la massima sensibilità verso tali esigenze culturali, assumendo parzialmente l'onere del finanziamento e dell'organizzazione di questo volume.

Esso certamente non sarà un momento isolato di collaborazine fra l'Università e l'Istituto Mecenate ma piuttosto rappresenta un punto di partenza ed un esempio da imitare.

Contributi determinanti alla realizzazione del volume sono dovuti al Dott. Giuseppe Di Biase, ricercatore di Istituzioni di Matematiche presso l'Università, che si è assunto l'impegno di una notevole mole di lavoro organizzativo e alla Prof.ssa Giuseppina Varone, docente di Matematica Applicata presso l'Istituto Mecenate, che ha rappresentato l'Istituto dal punto di vista scientifico.

Ringraziamenti particolari per la realizzazione di quest'opera vanno anche al Prof. Bruno Rizzi, ordinario di Analisi Matematica presso la Terza Università di Roma e Presidente Nazionale della Mathesis fino a tutto il 1993, al Prof. Romano Scozzafava, ordinario di Calcolo delle Probabilità e Statistica presso l'Università La Sapienza di Roma ed al Prof. Camillo Ciarlante, professore presso il Liceo Scientifico di Isernia ed organizzatore del Convegno di Isernia.

Franco Eugeni
professore di Geometria
Terza Università di Roma

Antonio Maturo
professore di Istituzioni di Matematiche
Università "G. D'Annunzio" di Chieti

Ratio Math. n. 7 - 1994
1 - 11

UNITÁ DIDATTICA:
UN METODO DI RAGGRUPPAMENTO – I VICINI RECIPROCI –

Carlo Antonelli
Istituto Tecnico Commerciale "R. De Sterlich" - Chieti Scalo

1. DALLA PROBLEMATICA ALLA METODOLOGIA DELL'ANALISI DEI GRUPPI.

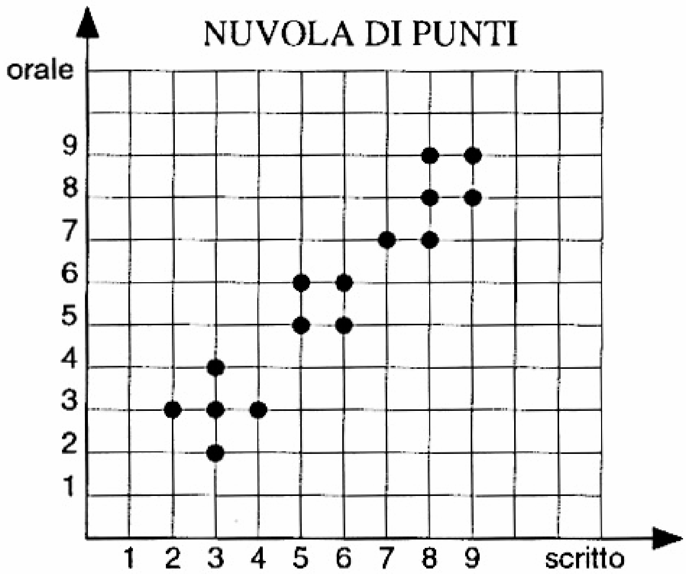
In una classe si vogliono individuare tre gruppi di alunni, il più possibile omogenei, rispetto al grado di preparazione in una certa materia.

Si rilevano, per ogni ragazzo, i voti riportati nello scritto e nell'orale in questa materia.

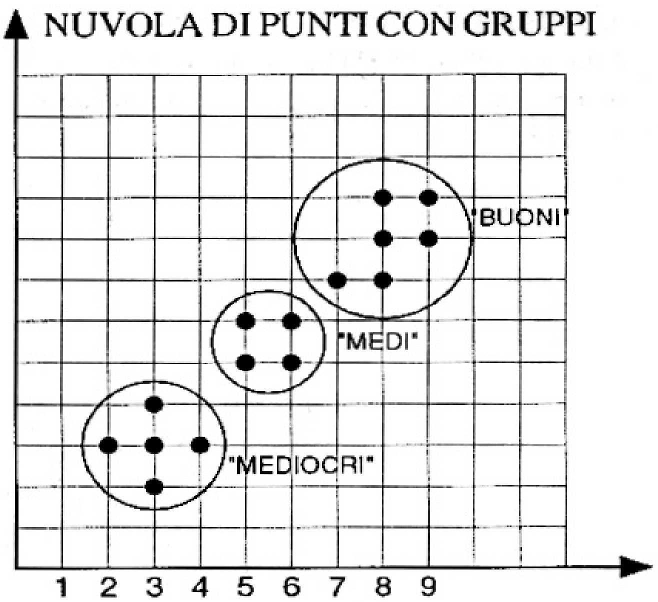
Si riportano le votazioni in un riferimento cartesiano, cosicché ogni alunno sarà individuato da un punto le cui coordinate sono rappresentate dai voti di quel ragazzo nello scritto (x) e nell'orale (y).

Nell'ipotesi in cui i voti rilevati sono i seguenti, avremo la nuvola di punti:

Alunno	Voto scritto	Voto orale
1	5	5
2	7	7
3	4	3
4	6	5
5	8	7
6	3	4
7	8	8
8	3	3
9	3	2
10	6	6
11	8	9
12	9	9
13	5	6
14	2	3
15	9	8

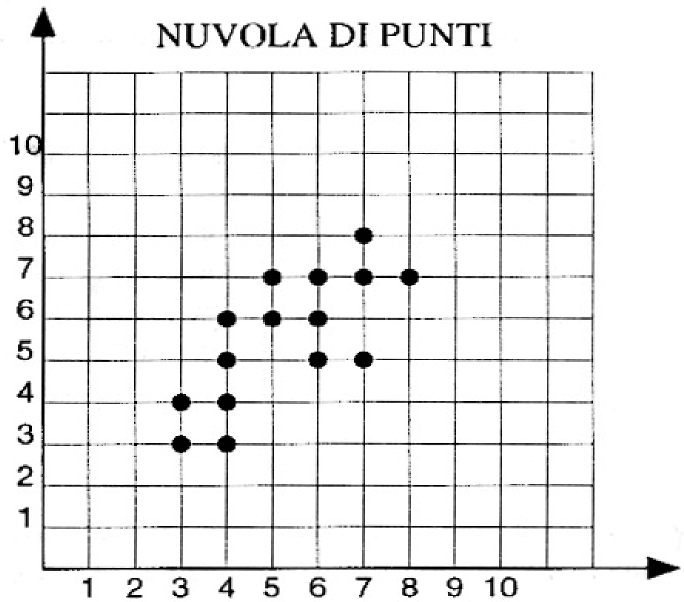


In questa situazione la sola analisi visiva della nuvola dei punti permette agevolmente e in maniera univoca, l'individuazione dei tre gruppi di alunni: gruppo dei "buoni", gruppo dei "medi", e gruppo dei "mediocri.



Ma situazioni del genere, in cui una popolazione risulta, rispetto ad una o più possibili variabili, "naturalmente" ben suddivisa in gruppi, sono abbastanza rare.

Nelle applicazioni pratiche è molto più facile incontrare situazioni, come quella riportata di seguito, in cui la formazione dei gruppi, per mezzo della sola analisi visiva, non è così facile ed univoca come nel caso precedente ma è lasciata all'arbitrarietà dell'individuo.



Si avverte l'esigenza, quindi, di usare delle metodologie statistiche che permettano, fissati alcuni presupposti che riguardano il tipo di distanza fra unità e la tecnica di raggruppamento, in maniera oggettiva e quindi ripetibile, l'individuazione dei gruppi.

Tali metodologie statistiche di classificazione che servono per raggruppare delle unità vanno sotto il nome di **ANALISI DEI GRUPPI o CLUSTER ANALISYS.**

2. INTRODUZIONE

Lo sviluppo dell'Analisi dei Gruppi ha avuto all'inizio degli anni 60 un grosso impulso grazie soprattutto a ricercatori di scienze naturali quali biologi e zoologi e da antropologi ed archeologi i quali avevano tutti l'esigenza di mettere a punto metodi matematici che permettessero di raggruppare o classificare insiemi di unità.

Queste tecniche di raggruppamento hanno subito trovato un largo campo di applicazione in tutte quelle situazioni dove sorge l'esigenza di raggruppare popolazioni in gruppi (clusters) di unità il più possibile omogenee come per esempio nelle scienze sociali (**individuazione di classi sociali e comportamentali**), nelle ricerche di mercato (**individuazione di stili di vita, segmentazione del mercato**) e nella teoria dei campioni (**individuazione degli strati nei campioni stratificati**).

L'enorme potenziamento dei calcolatori in termini di velocità di calcolo e capacità di memoria ha permesso l'effettivo uso di tali procedure in situazioni reali cioé con popolazioni di migliaia di unità e decine di variabili.

3. TIPOLOGIE DEI METODI DI RAGGRUPPAMENTO

I metodi di classificazione si suddividono in due gruppi: – **metodi scissori** mediante i quali la formazione dei gruppi avviene suddividendo la popolazione di partenza in classi sempre più numerose;

– **metodi aggregativi** mediante i quali la formazione dei gruppi avviene aggregando per passi successivi le singole unità di partenza in gruppi di numerosità crescente.

Si parte considerando ogni unità statistica gruppo a se e per aggregazioni successive, intorno a queste unità iniziali, si formano passo dopo passo numeri decrescenti di gruppi fino a riaggregare le unità statistiche in un solo gruppo che rappresenta la popolazione di partenza.

Il metodo scelto per questa unità didattica è l'algoritmo dei "**VICINI RECIPROCI**" metodo aggregativo che per la sua "velocità" di aggregazione rispetto agli altri algoritmi viene molto usato nelle ricerche sociali dove si ha a che fare con popolazioni molto numerose.

Il principio del metodo dei Vicini Reciproci risale a Mc Quitty (1966), la sua messa a punto è dovuta a C. de Rham (1980) e la sua prima sperimentazione operativa a Benzecri (1982).

Il metodo è stato scelto per questa unità didattica grazie ella seguenti sue caratteristiche:
– è agevole e veloce nell'uso;
– richiede allo studente come prerequisiti, un limitato numero di conoscenze e abilità;
– fa uso anche di strumenti e concetti propri di altre discipline (coordinate, punti, distanze);
– è esaustivo sia per quanto riguarda i concetti base dell'analisi dei gruppi sia per quelli di una completa indagine statistica;
– da la possibilità di "complicare" la procedura introducendo concetti e strumenti più "sofisticati" (standardizzazione delle variabili, particolari distanze), o di inserirlo in un contesto più ampio (teoria dei campioni), in modo da poterlo trattare proficuamente a vari livelli ed in diversi momenti di un corso di statistica.

4. ALGORITMO DEI "VICINI RECIPROCI"

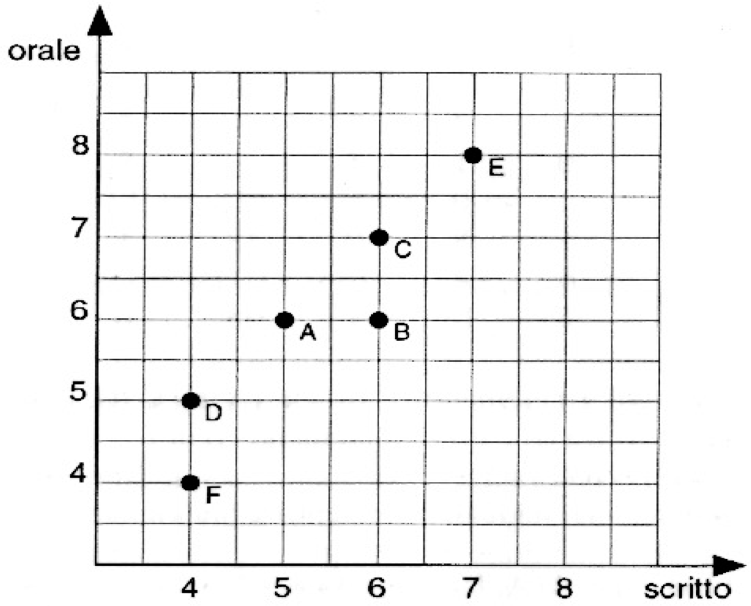
Esempio di applicazione dell'algoritmo aggregativo dei "Vicini Reciproci" su una popolazione di 6 unità statistiche caratterizzate da due variabili quantitative.

Obiettivo: individuare in un collettivo di 6 alunni tre gruppi di ragazzi omogenei rispetto al grado di preparazione in matematica. Come variabili esplicative del grado di preparazione vengono prese le votazioni riportate dai 6 alunni nello scritto e nell'orale di matematica.

1ª fase: rilevazione dei dati.
Insieme delle osservazioni $I = \{A, B, C, D, E, F\}$

Alunno	Voto scritto	Voto orale
A	5	6
B	6	6
C	6	7
D	4	5
E	7	8
F	4	4

2ª fase: rappresentazione della nuvola di punti



3ª fase: calcolo delle (6x5) / 2 = 15 distanze d (i,j). Come distanza è stata scelta il quadrato delle distanza euclidea $d = [(x_i + x_j)^2 + (y_i + y_j)^2]$ e costruzione della matrice delle distanze.

$$d(A, B) = [(5-6)^2 + (6-6)^2] = 1$$
$$d(A, C) = [(5-6)^2 + (6-7)^2] = 2$$
$$d(A, D) = [(5-4)^2 + (6-5)^2] = 2$$
$$d(A, E) = [(5-7)^2 + (6-8)^2] = 8$$
$$d(A, F) = [(5-4)^2 + (6-4)^2] = 5$$
$$d(B, C) = [(6-6)^2 + (6-7)^2] = 1$$
$$d(B, D) = [(6-4)^2 + (6-5)^2] = 5$$
$$d(B, E) = [(6-7)^2 + (6-8)^2] = 5$$
$$d(B, F) = [(6-4)^2 + (6-4)^2] = 8$$
$$d(C, D) = [(6-4)^2 + (7-5)^2] = 8$$
$$d(C, E) = [(6-7)^2 + (7-8)^2] = 2$$
$$d(C, F) = [(6-4)^2 + (7-4)^2] = 13$$
$$d(D, E) = [(4-7)^2 + (5-8)^2] = 18$$
$$d(D, F) = [(4-4)^2 + (5-4)^2] = 1$$
$$d(E, F) = [(7-4)^2 + (8-4)^2] = 25$$

MATRICE DELLE DISTANZE

	A	B	C	D	E	F
A	0	1	2	2	8	5
B	1	0	1	5	5	8
C	2	1	0	8	2	13
D	2	5	8	0	18	1
E	8	5	2	18	0	25
F	5	8	13	1	25	0

4ª fase: ricerca per tutti gli elementi "i" di I = {A, B, C, D, E, F} della minima distanza min d (i) fra (i) e gli altri elementi (j) di I.

Min d (i) = inf [d(i, j) / jEI, j $\neq$ i].

Ricerca per ogni i del v (i), cioé dell'elemento il più vicino ad i.

min d (A) = 1 = d (A, B) $\Rightarrow$ v (A) = B
min d (B) = 1 = d (B, A), d (B, C) $\Rightarrow$ v (B) = A e C
min d (C) = 1 = d (C, B) $\Rightarrow$ v (C) = B
min d (D) = 1 = d (D, F) $\Rightarrow$ v (D) = F
min d (E) = 2 = d (E, C) $\Rightarrow$ v (E) = C
min d (F) = 1 = d (F, D) $\Rightarrow$ v (F) = D

Schematizzando avremo:

i:	A	B	C	D	E	F
min d (i):	1	1	1	1	2	1
v (i):	B	A,C	B	F	C	D

5ª fase: individuazione delle coppie dei Vicini Reciproci che andranno a formare l'insieme M.

Definizione: due elementi i, j si definiscono Vicini Reciproci se min d (i) = min d (j), e se ogni elemento della coppia è l'elemento più vicino dell'altro elemento.

Deve verificarsi che i = v (j) e j = v (i).

Così gli elementi A e B sono Vicini Reciproci perché A = v(B) e B = v(A), A è l'elemento il più vicino a B e B è l'elemento il più vicino ad A.

Le coppie di Vicini Reciproci saranno quindi:

(A e B) in quanto v (A) = B e v (B) = A;
(D e F) in quanto v (D) = F e v (F) = D.

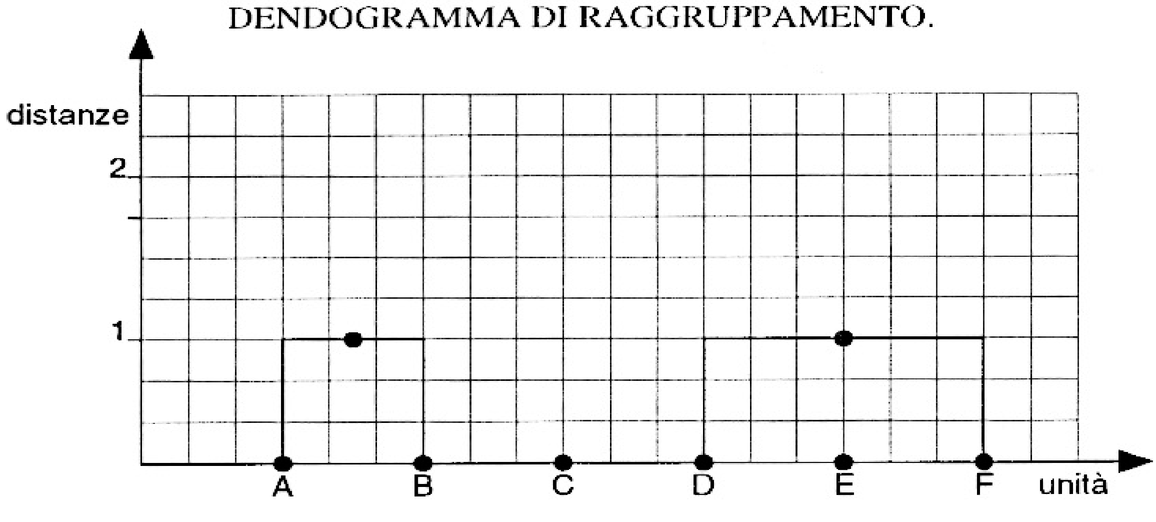
Gli elementi C ed E non Formano una coppia di V. R. in quanto:
v(C) = B e v(E) = C.

L'insieme M delle coppie dei V. R. sarà M = {(A, B), (D, F)}.

La prima iterazione è terminata ed ha prodotto i seguenti 4 gruppi:
I = {(A, B), C, (D, F), E}.

Si può fornire una rappresentazione grafica (DENDOGRAMMA) di questa procedura che poi viene aggiornata ad ogni nuova iterazione.
Sull'asse delle ascisse vengono riportate le unità statistiche, sull'asse delle ordinate le distanze a cui avvengono le aggregazioni.

DENDOGRAMMA DI RAGGRUPPAMENTO.



Al livello di distanza 1 l'algoritmo aggrega le unità A e B e le unità D e F in quanto coppie di Vicini Reciproci. I nodi a livello 1 indicano la formazione dei gruppi.

Tornando alla nuvola di punti la prima iterazione ha prodotto 4 gruppi.

II ITERAZIONE

A questo punto si itera la procedura con la novità che le distanze da calcolare non sono più tra elementi ma tra gruppi ed elementi e tra gruppo e gruppo.

Come distanza tra gruppo ed elemento prendiamo la distanza media. La distanza tra l'unità C e la classe (A, B) è data dalla media aritmetica delle distanze tra C e le singole unità delle classe (A, B):

d [(C, (A, B)] = [d(C, A) + d (C, B)] / 2. Per quanto riguarda la distanza fra le classi prendiamo la distanza media interclasse, cosicché la distanza fra i gruppi (A, B) e (D, F) sarà data da:

d [(A, B), (D, F)] = [d(A, D) + d (A, F) + d (B, D) + d (B, F)] / 4.

Calcoliamo le nuove distanze:

d [(A, B), C] = [d(A, C) + d(B, C)] / 2 = (2 + 1) / 2 = 1.5
d [(A, B), E] = [d(A, E) + d(B, E)] / 2 = (8, 5) / 2 = 6.5
d [(A, B), (D, F)] = [d(A, D) + d(A, F) + d(B, D) + d(B, F)] / 4
= (2 + 5 + 5 + 8) / 4 = 5
d [(C, (D, F)] = [d(C, D) + d(C, F)] / 2 = (8 + 13) / 2 = 10.5
d [(D, F), E] = [d(D, E) + d(F, E)] / 2 = (18 + 25) / 2 = 21.5

e la nuova matrice delle distanze

	(A, B)	**C**	**(D, F)**	**E**
(A, B)	0	1.5	5	6.5
C	1.5	0	10.5	2
(D, F)	5	10.5	0	21.5
E	6.5	2	21.5	0

Cerchiamo per tutti gli elementi di "i" di I = {(A, B), C, (D, F), E} la minima distanza min d (i) fra (i) e gli altri elementi (j) di I. Ricerchiamo per ogni i il v(i) cioé l'elemento il più vicino ad i.

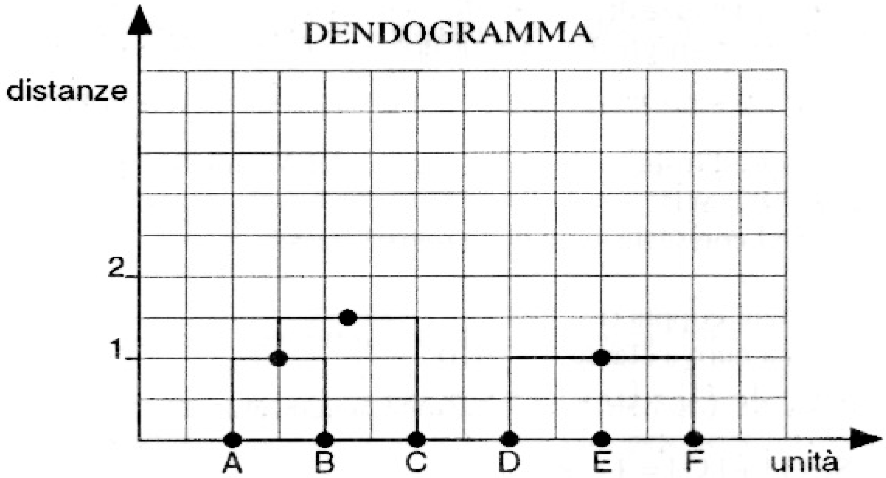
Schematizzando avremo:

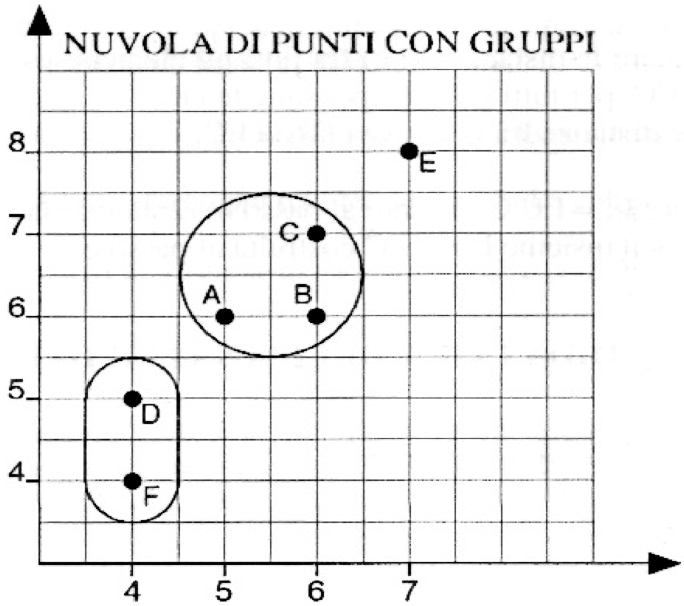
i:	(A, B)	C	(D, F)	E
min d(i):	1.5	1.5	5	2
v(i):	C	(A, B)	(A, B)	C

Quindi gli elementi (A, B) e C sono Vicini Reciproci in quanto v(A, B) = C e v(C) = (A, B).

La seconda iterazione è terminata ed ha prodotto i seguenti tre gruppi: (A, B, C), (D, F), (E).

Possiamo aggiornare il nostro dendogramma che ci descrive tutte le fasi della formazione dei gruppi.





Poiché il nostro obiettivo era quello di suddividere la popolazione di partenza in tre gruppi interrompiamo la procedura che, se reiterata, porterebbe alla costituzione di un unico gruppo, la popolazione di partenza.

La scelta del numero ottimale di gruppi, se non viene stabilita a priori, come nel nostro caso, viene di solito effettuata analizzando il dendogramma o con l'ausilio di appropriati indici.

È possibile formalizzare in "passi" l'Algoritmo dei Vicini Reciproci sia per una trattazione e presentazione di tipo formale sia per tradurre l'algoritmo in procedura informatica.

Passo a) Costruzione della matrice delle distanze.
1) Calcolare le distanze d(i, j) per tutti i, jEI, i ≠ j.
2) Ricercare, per tutti gli iEI, la min d(i) e i v(i) (l'elemento il più vicino ad i).

Passo b) Trovare l'insieme M delle coppie dei Vicini Reciproci M={(i,j) i,jEI, i≠j, i=v(j), j=v(i)}.
Definire C: = Ø con C insieme delle nuove classi.

Passo c) Per ogni coppia (i,j) EM
1) Formare una nuova classe c = (iUj).
2)Trovare I : = I - {i} - {j} e C; = C + {c}.

Passo d) Se | I | + | C | = 1 : stop.

Passo e) Calcolare le distanze d(i,c) fra tutti gli elementi esistenti iEI e le nuove classi cEC, per tutti gli iEI e per tutte le cEC.
2) Calcolare le distanze fra le nuove classi cEC.

Passo f) Trovare I : = I U C. Andare al passo b). 2ª iterazione sulla matrice delle distanze dell'insieme I = I U C costruita al passo e).

5. VALENZA DIDATTICA DELL'ANALISI DEI GRUPPI

Riguardo alle valenze didattiche di tale procedura possiamo dire che: la metodologia dell'ANALISI DEI GRUPPI è una indagine statistica completa in quanto riassume in se tutte le fasi di una ricerca statistica, dalla raccolta dei dati alla tabulazione, dalla rappresentazione grafica alla eleborazione dei dati e all'analisi dei risultati. Pertanto essa può essere trattata proficuamente al termine di un corso di statistica descrittiva.

Si può calibrare il grado di difficoltà, rapportandolo al livello del corso. Infatti partendo da un livello minimale in cui sono richiesti all'alunno pochi prerequisiti, si possono introdurre nella procedura di raggruppamento delle "migliorie" come la standardizzazione delle variabili, l'uso di altri tipi di distanza legate anche alla tipologia delle variabili (centro di gravità, varianza interclasse) o indici per la scelta della partizione migliore.

BIBLIOGRAFIA

[1] A. Bellacicco *Metodologia e tecnica della classificazione matematica* - La goliardica- Roma (1975)

[2] J. P. Benzecri *Le Cahiers de l'analyse des Donnée* - Vol. VII (1982).

[3] S. Zani *Seminario: Due temi di analisi statistica multivariata* - Cleup Bressanone (1978)

[4] S. ZANI *L'analisi classificatoria - Contributi metodo-logici ed impiego per l'individuazione di aree omogenee.* Istituto di statistica, Università degli studi di Parma (1977)

[5] G. LETI *Statistica descrittiva* – Il Mulino Bologna (1983)

[6] M. JAMBU *Classification automatique pour l'anlyse des données.* Dunod Paris (1878)

[7] S. BOLASCO - R. Coppi *Analisi dei dati e sue applicazioni nella ricerca economica* - ed. Klim Roma (1983)

[8] AA.VV. *Introduzione allo SPAD* – Centro di calcolo interfacoltà (1984)

[9] A. RIZZI *Analisi dei gruppi (cluster analysis)* - La goliardica (1981)

[10] G. MARBACH *Le ricerche di mercato* – UTET (1982)