I modelli di analisi statistica multidimensionale dei dati:

La Cluster Analysis Non Gerarchica

Obiettivi dell'unità didattica

- Comprendere i concetti generali dei principali metodi di classificazione non gerarchica
- Mostrare i comandi principali su Tanagra del metodo delle k-medie

4.7 Cluster Analysis Non Gerarchica

- **Definizione della CA non gerarchica**
- **Metodo delle Nubi Dinamiche**
- **Metodo delle k-medie**
- **Esempio: il congresso americano**

Classificazione non gerarchica

Richiede la determinazione a priori del numero di classi che definiscono la partizione.

L'algoritmo è convergente e il numero di iterazioni richieste è generalmente limitato: ciò rende questo metodo applicabile anche a grossi insiemi di dati

La soluzione ottenuta non rappresenta la soluzione ottimale ma solo una delle tante possibili, ottenuta avendo determinato a priori quel numero di classi e avendo scelto quelle unità iniziali.

Principali metodi:

- Metodo delle nubi dinamiche (Forgy, 1965)
- Metodo delle k-medie (k-means, Mac Queen, 1967)

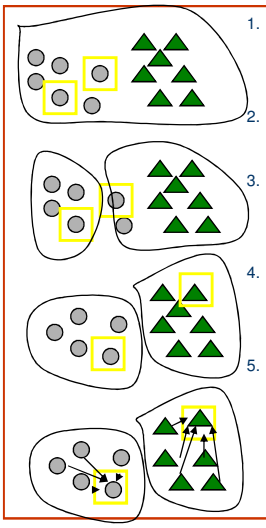
Metodo delle nubi dinamiche

Perviene ad una partizione di E attraverso un procedimento iterativo che, a partire da una soluzione iniziale arbitraria, la "migliora" fino a pervenire a quella "ottima", definita tale in base ad un determinato criterio.

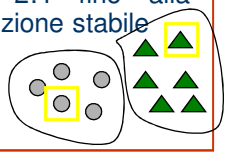
L'algoritmo richiede generalmente poche iterazioni e consente il trattamento di grossi insiemi di dati ma:

- Il numero delle classi deve essere fissato a priori
- La soluzione dipende dalle assegnazioni (nuclei) iniziali
- Non esiste un unico criterio ottimale, per cui la partizione ottimale è una tra le tante possibili partizioni

L'algoritmo del metodo delle nubi dinamiche



1. Definizione del numero delle classi e selezione delle unità rappresentative (nuclei di ciascuna classe
2. Calcolo delle distanze tra i nuclei e le unità attribuite al punto seme più vicino
3. Definizione dei nuovi nuclei per ciascuna classe considerando le unità più vicine al centroide del gruppo
4. Calcolo delle distanze tra i nuovi nuclei e le unità statistiche allocate ai gruppi più vicini
5. Ripetizione dei passi 2:4 fino alla convergenza ad una soluzione stabile



Metodo delle k-medie: algoritmo

- Si assumono come nuclei i primi K individui, si allocano via via le n-K unità e ad ogni assegnazione si calcola subito il centroide del gruppo che si è modificato: in tal modo si accelera il miglioramento della classificazione.
- Al termine di questa prima assegnazione si riparte dai centroidi così calcolati e si riassegnano le unità via via al gruppo più vicino.
- Regola di stop: come per il metodo delle nubi dinamiche.

Punti di debolezza del metodo delle k-medie

- Spesso termina su un ottimo locale. L'ottimo globale può essere trovato usando tecniche alternative come: deterministic annealing e genetic algorithm.
- Può essere applicato solo quando il tipo di dato permette di definire la media (che serve per determinare i centroidi del cluster)→Problemi con dati categorici.
- Bisogna specificare in anticipo k, il numero di cluster.

Esempio: i membri del congresso

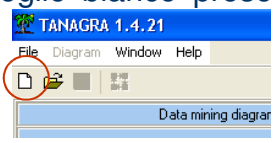
Come caso studio per la cluster non gerarchica si prende in considerazione il famoso dataset “vote”, che analizza attraverso il voto dei rappresentati della Camera dei Rappresentanti l’appartenenza ad uno schieramento o all’altro.

Il dataset, anche se non molto recente, ci permette di elaborare 17 votazioni dei membri (435 individui) del congresso americano.

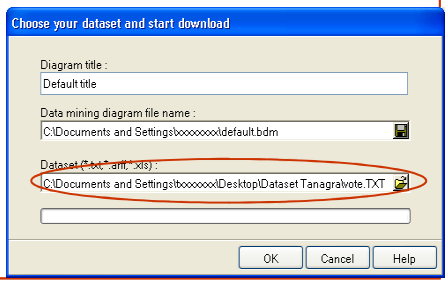
La variabile “class” indica l’appartenenza o al partito democratico o a quello repubblicano e non entrerà a far parte dell’analisi, ma la si utilizzerà per confrontare il risultato finale della CA.

Esempio: aprire un file .txt

Il file “vote” è un file di testo, quindi, per aprirlo è sufficiente aprire Tanagra e cliccare sull’immagine del foglio bianco presente sotto la voce del menù File.

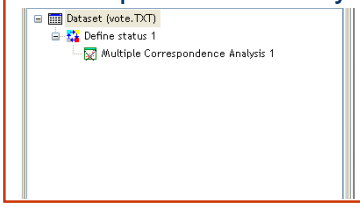


Si aprirà così un finestra da cui selezionare (nello spazio cerchiato in rosso sulla destra) il file d’interesse.



Esempio: sintetizzare le variabili qualitative

Una volta aperto il file si vedrà che le variabili sono tutte discrete, di conseguenza sarà necessario sintetizzare queste variabili qualitative in variabili di sintesi quantitative attraverso l'ACM. Si aggiunga, quindi, prima il comando Define Status (come input inserire tutte le variabili, tranne la variabile "class) e, successivamente, il componente Multiple Correspondence Analysis.

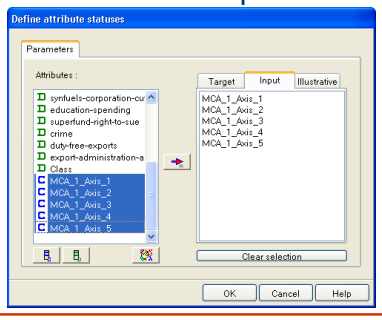


Eigen values

Axis	Eigen value	% explained	Histogram	% cumulated
1	0,472217	23,61%		23,61%
2	0,269415	13,47%		37,08%
3	0,096981	4,85%		41,93%
4	0,090299	4,51%		46,45%
5	0,075743	3,79%		50,23%

Esempio: inizio della CA non gerarchica

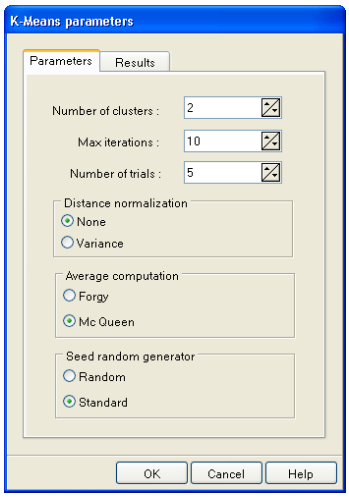
I primi 5 assi fattoriali raggiungono il 50% dell'informazione disponibile. Si possono utilizzare questi assi come descrittori per il metodo delle k-medie. Si aggiunga nuovamente il comando Define Status e come input si inseriscano gli assi fattoriali.



Esempio: il componente K-Means

Si aggiunga il componente *K-Means* che si trova sotto la macrovoce Factorial Analysis, si clicchi con il tasto destro del mouse sul nuovo componente e si lascino le impostazioni mostrate nella figura sulla destra.

Settiamo 2 cluster perché sono due i partiti presenti nella Camera dei Rappresentanti. Quindi, cliccando con il tasto destro del mouse su K-Means, si selezioni View.



Esempio: composizione dei due cluster

Il primo cluster è composto da 197 individui, il secondo da 238. L'inerzia spiegata è quasi pari al 40%. Per capire in che modo sono stati strutturati i cluster, si aggiunga nuovamente un Define Status, si inserisce come Target la nuova "variabile" Cluster_Kmeans_1 e come Input tutte le variabili originali (questa volta anche la variabile "class"), mentre gli assi non devono essere inseriti. Si aggiunga, quindi, il componente *Group Characterization* in Statistics. Infine, cliccando con il tasto destro sull'ultimo componente aggiunto, si selezioni View.

Esempio: confronto con i due partiti

Confrontando i due cluster con i due partiti si può verificare che nel primo cluster si trova il 79% dei repubblicani, mentre nel secondo il 95% dei democratici. Ciò sta a dimostrare come la composizione dei due cluster sia molto vicina alla reale affiliazione politica dei membri della Camera.

Cluster_KMeans_1=c_kmeans_1			
Examples		[45,3 %] 197	
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
Discrete attributes : [Recall] Accuracy			
el-salvador-aid='y'	18,1	[89,6 %] 96,4 %	48,7 %
aid-to-nicaraguan-contras='n'	17,1	[94,4 %] 85,3 %	40,9 %
physician-fee-freeze='y'	17,0	[94,4 %] 84,8 %	40,7 %
mx-missile='n'	16,1	[85,9 %] 89,8 %	47,4 %
Class='republican'	15,8	[92,9 %] 79,2 %	38,6 %
adoption-of-the-budget-re='n'	15,7	[91,8 %] 79,7 %	39,3 %
education-spending='y'	14,7	[88,9 %] 77,2 %	39,3 %
crime='y'	14,5	[75,4 %] 94,9 %	57,0 %
anti-satellite-test-ban='n'	14,4	[85,7 %] 79,2 %	41,8 %

Cluster_KMeans_1=c_kmeans_2			
Examples		[54,7 %] 238	
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
Discrete attributes : [Recall] Accuracy			
el-salvador-aid='n'	17,8	[99,0 %] 86,6 %	47,8 %
aid-to-nicaraguan-contras='y'	17,5	[92,1 %] 93,7 %	55,6 %
physician-fee-freeze='n'	16,5	[89,1 %] 92,4 %	56,8 %
Class='democrat'	15,8	[84,6 %] 95,0 %	61,4 %
adoption-of-the-budget-re='y'	15,3	[85,8 %] 91,2 %	58,2 %
mx-missile='y'	14,8	[91,8 %] 79,8 %	47,6 %
crime='n'	14,4	[97,6 %] 69,7 %	39,1 %
education-spending='n'	14,2	[86,3 %] 84,5 %	53,6 %
anti-satellite-test-ban='y'	14,2	[85,4 %] 85,7 %	54,9 %